Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: (26. June 2024)

Kursus navn og nr.: **Introduktion til Statistik (02402)**

Varighed: 4 timer

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

_____ _____ _____

(studienummer) (underskrift) (bord nr.)

Opgavesættet består af 30 spørgsmål af "multiple choice" typen, som er fordelt på 12 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" siderne på eksamen.dtu.dk.

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

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

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

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

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

Eksamenssættet består af 26 sider.

Fortsæt på side 2

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

Opgave I

To studerende tæller antallet af biler, der kører forbi på forskellige vejstrækninger. De antager, at antallet af biler, der passerer i bestemte tidsintervaller, følger Poisson-fordelinger. På den første vej (vej 1) antager de, at det forventede antal biler der passerer forbi er $\lambda_1 = 10/\text{time}$, mens de på den anden vej (vej 2) antager, at det forventede antal biler der passerer forbi er $\lambda_2 = 15/\text{time}$.

De definerer nu to stokastiske variable

- X_1: antal biler, der kører forbi på vej 1 på 15 minutter
- X_2: antal biler, der kører forbi på vej 2 på 10 minutter.

Du kan antage, at X_1 og X_2 er uafhængige.

Spørgsmål I.1 (1)

Hvad er sandsynligheden $P(X_1 = 10)$?

1 ☐ 0.125

2 ☐ 0.417

3 ☐ 0.000216

4 ☐ 0.875

5 ☐ 0.583

Fortsæt på side 3

Spørgsmål I.2 (2)

Hvilket af følgende udsagn om de forventede værdier af de to stokastiske variable er korrekt?

1 ☐ $\frac{E[X_1]}{E[X_2]} = 1.5$

2 ☐ $\frac{E[X_1]}{E[X_2]} = \frac{2}{3}$

3 ☐ $\frac{E[X_1]}{E[X_2]} = \frac{1}{3}$

4 ☐ $\frac{E[X_1]}{E[X_2]} = 3$

5 ☐ $\frac{E[X_1]}{E[X_2]} = 1$

Spørgsmål I.3 (3)

Hvad er sandsynligheden for, at der går mere end 2 minutter i mellem to på hinanden følgende biler, på vej 2?

1 ☐ 0.5

2 ☐ 0.184

3 ☐ 0.607

4 ☐ 0.368

5 ☐ 0.816

Fortsæt på side 4

Opgave II

En gård lavede en undersøgelse, hvor 225 kyllinger blev tilfældigt opdelt i 3 behandlingsgrupper på hver 75 dyr. Hver gruppe blev fodret med foder fra forskellige foderproducenter i en periode. For hver kylling blev ændring i vægten over perioden målt, og det endelige datasæt består derfor af 225 observationer af vægtændringer. Formålet med undersøgelsen er at afgøre, om der er statistisk evidens for forskel i middelvægtændring for mindst én af grupperne. Det kan antages, at variansen er ens for alle behandlingsgrupper.

Spørgsmål II.1 (4)

Hvilken type statistisk analyse er bedst egnet til dette?

1 ☐ Multipel lineær regressionsanalyse

2 ☐ Test for uafhængighed i en $r \times c$ frekvenstabel (antalstabel)

3 ☐ Parret t-tests

4 ☐ Envejs variansanalyse

5 ☐ t-tests

Fortsæt på side 5

Opgave III

Ingeniørerne i en international lufthavn har lavet en undersøgelse, hvor de har taget tid på 40 tilfældigt udvalgte sikkerhedstjek. Den gennemsnitlige varighed af sikkerhedskontrollen inkluderet i undersøgelsen var 34.66 sekunder, og stikprøvens standardafvigelse var 10.12 sekunder, det antages, at tiderne er normalfordelt.

Spørgsmål III.1 (5)

Baseret på undersøgelsen, hvad er 99% konfidensintervallet for den gennemsnitlige varighed af sikkerhedskontrollen?

1 ☐ $[7.26; 62.06]$

2 ☐ $[14.19; 55.13]$

3 ☐ $[30.33; 38.99]$

4 ☐ $[31.42; 37.90]$

5 ☐ $[33.06; 36.26]$

Spørgsmål III.2 (6)

Hvad er p-værdien for den sædvanlige test af nul-hypotesen $H_0 : \mu = 30$ mod en tosidet alternativ hypotese?

1 ☐ 0.30%

2 ☐ 0.59%

3 ☐ 4.72%

4 ☐ 94.23%

5 ☐ 99.70%

Fortsæt på side 6

Opgave IV

En virksomhed ønsker at estimere omkostningerne ved at producere solpaneler. Derfor har ingeniørerne designet et eksperiment for at vurdere omkostningerne (costs) ved at producere batches af forskellig størrelse (batch size) og sikre, at observationerne er helt uafhængige. Resultaterne er angivet i tabellen nedenfor.

Batch size (units)	50	100	150	200	250	300	350	400	450	500
Costs (M DKK)	2.33	4.21	6.01	7.51	8.46	8.93	9.45	10.70	10.55	10.74

Data kan læses i R ved hjælp af nedenstående kode:

```
Batch<-1:10 * 50
Costs<-c(2.33,4.21,6.01,7.51,8.46,8.93,9.45,10.70,10.55,10.74)
```

Ingeniørerne mener i første omgang, at data kan beskrives ved en lineær model på formen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i \in \{1, ..., 10\},$$

hvor fejlene, ϵ_i, er uafhængige og identisk fordelt (i.i.d.) med en $N(0, \sigma^2)$-fordeling. I modellen er responsvariablen omkostningerne (costs) (i M DKK) og den forklarende variabel er batchstørrelsen (i enheder). Ingeniørerne estimerer derfor en lineær regressionsmodel ved hjælp af mindste kvadraters metode.

Spørgsmål IV.1 (7)

Hvilken andel af den total variation i omkostningerne forklares af regressionsmodellen?

1 ☐ 89.4%

2 ☐ 90.6%

3 ☐ 93.9%

4 ☐ 95.2%

5 ☐ 96.9%

Fortsæt på side 7

Spørgsmål IV.2 (8)

Hvad er 99% konfidensintervallet for hældningen, β_1?

1 ☐ $[0.011, 0.025]$

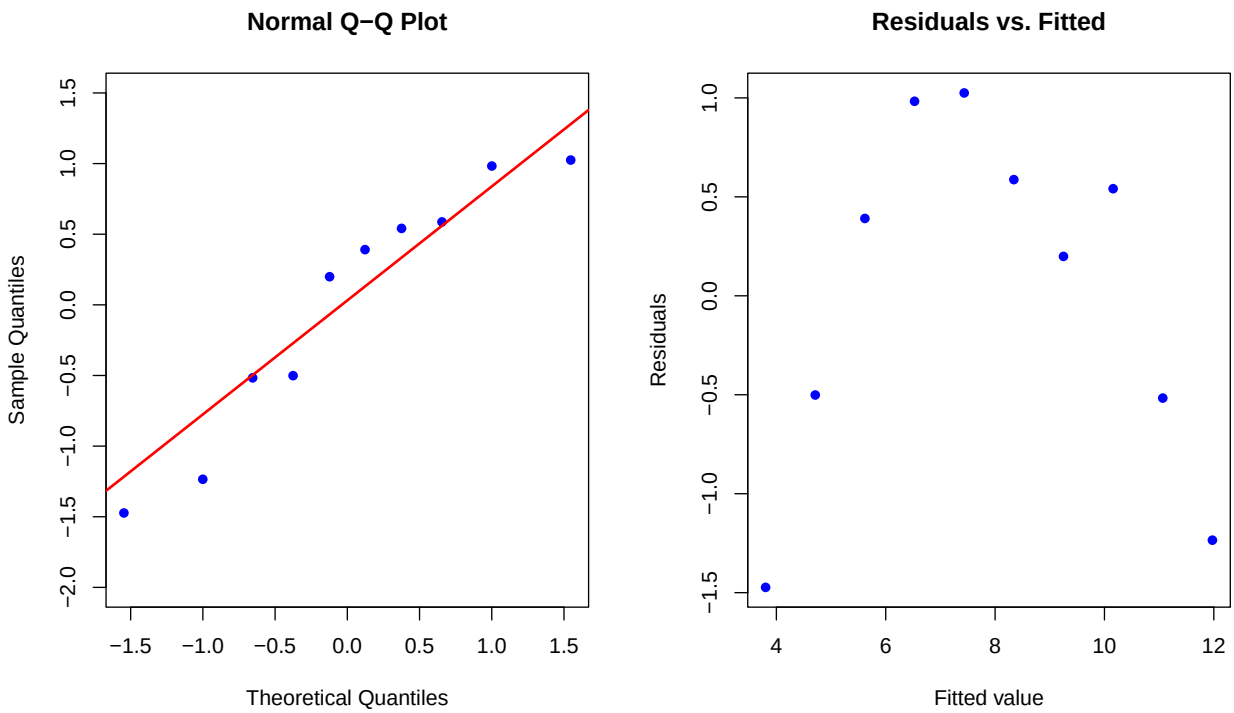
2 ☐ $[0.013, 0.023]$

3 ☐ $[0.016, 0.020]$

4 ☐ $[0.742, 5.049]$

5 ☐ $[1.415, 4.375]$

For at validere modellen laver ingeniørerne to diagnostiske plots: Et normalt Q-Q plot og et plot af residualerne mod de fittede værdier. De to plots ses nedenfor:



Fortsæt på side 8

Spørgsmål IV.3 (9)

Hvilket af følgende udsagn om modellens gyldighed er mest korrekt?

1 ☐ De diagnostiske plots indikerer ikke nogen overtrædelser af modellens antagelser

2 ☐ Det normale Q-Q plot indikerer, at residualerne ikke er normalfordelte

3 ☐ Plottet med residualer vs. fittede værdier indikerer, at residualerne ikke er normalfordelte

4 ☐ Plottet med residualer vs. fittede værdier indikerer, at residualerne og de fittede værdier ikke er uafhængige

5 ☐ Modelantagelserne må være opfyldt, da R^2-værdien er høj.

Fortsæt på side 9

Opgave V

Antallet af personer, der bor på danske kollegier i 2023, oplyses af Danmarks Statistik. Her fokuserer vi kun på udvalgte aldersgrupper.

	mænd	kvinder	Total
18-24	14048	14128	28176
25-29	8215	6028	14243
30-39	2735	1397	4132
Total	24998	21553	46551

Spørgsmål V.1 (10)

Hvor stor en andel af beboerne på danske kollegier er mænd (blandt de 18-39 årige)?

1 ☐ 0.463

2 ☐ 0.500

3 ☐ 0.521

4 ☐ 0.537

5 ☐ 0.409

Spørgsmål V.2 (11)

Vi vil gerne vide om, andelen af mænd er signifikant forskellig i de forskellige aldersgrupper (vi bruger et 5% signifikansniveau). Hvilket af de følgende udsagn er korrekt?

1 ☐ Vi skal bruge en parret t-test med $\alpha = 0.05$ for at teste, om der er en signifikant forskel mellem aldersgrupperne. Resultatet er, at der ER en signifikant forskel.

2 ☐ Vi skal bruge en uparret t-test med $\alpha = 0.025$ for at teste, om der er en signifikant forskel mellem aldersgrupperne. Resultatet er, at der ER en signifikant forskel.

3 ☐ Vi skal bruge en uparret t-test med $\alpha = 0.05$ for at teste, om der er en signifikant forskel mellem aldersgrupperne. Resultatet er, at der IKKE er en signifikant forskel.

4 ☐ Vi skal bruge en χ^2-test med 6 frihedsgrader for at teste, om der er en signifikant forskel mellem aldersgrupperne. Resultatet er, at der ER en signifikant forskel.

5 ☐ Vi skal bruge en χ^2-test med 2 frihedsgrader for at teste, om der er en signifikant forskel mellem aldersgrupperne. Resultatet er, at der ER en signifikant forskel.

Spørgsmål V.3 (12)

Under den hypotese, at fordelingen mellem mænd og kvinder er den samme i alle aldersgrupper, hvad er da det forventede antal mænd i aldersgruppen 18-24 der bor på kollegier (som kan sammenholdes med tabellen ovenfor og bruges til at beregne den relevante teststørrelse)?

1 ☐ 14088

2 ☐ 15131

3 ☐ 15517

4 ☐ 14048

5 ☐ 7759

Spørgsmål V.4 (13)

Vi ser nu på de 18-24 årige. Hvilket af de følgende udsagn er korrekt (i de svarmuligheder hvor det er relevant bruges signifikansniveau $\alpha = 0.05$)?

1 ☐ Blandt de 18-24 årige er andelen af mænd IKKE signifikant forskellig fra 0.5, da det estimerede 95% konfidensinterval for andelen af mænd i denne aldersgruppe er [0.493, 0.504]

2 ☐ Blandt de 18-24 årige er andelen af mænd præcis 0.5.

3 ☐ Blandt de 18-24 årige er andelen af mænd signifikant forskellig fra 0.5, da det estimerede 95% konfidensinterval for andelen af mænd i denne aldersgruppe er [0.501, 0.533]

4 ☐ Blandt de 18-24 årige er andelen af mænd IKKE signifikant forskellig fra 0.5, da det estimerede 95% konfidensinterval for andelen af mænd i denne aldersgruppe er [0.501, 0.533]

5 ☐ Blandt de 18-24 årige er andelen af mænd signifikant forskellig fra 0.5, da det estimerede 95% konfidensinterval for andelen af mænd i denne aldersgruppe er [0.532, 0.542]

Fortsæt på side 11

Spørgsmål V.5 (14)

Under antagelse af uafhængighed mellem individer, hvad er sandsynligheden for, at 100 eller flere kvinder bor på et kollegie med i alt 190 beboere, hvis sandsynligheden for, at en individuel beboer er en kvinde er lig med 0.45?

1 ☐ Sandsynligheden er 0.021

2 ☐ Sandsynligheden er 0.015

3 ☐ Sandsynligheden er 0.50

4 ☐ Sandsynligheden er 0.45

5 ☐ Sandsynligheden er 0.985

Fortsæt på side 12

Opgave VI

En simpel rov-byttedyr model er Lotka-Volterra-modellen

$$\frac{dx}{dt} = \alpha x - \beta xy$$
$$\frac{dy}{dt} = \delta xy - \gamma y,$$

hvor x er størrelsen af byttedyrbestanden og y er størrelsen af rovdyrbestanden. Ligningen giver mulighed for en bevægelseskonstant (dvs. en størrelse der forbliver konstant gennem tiden for givne begyndelsesbetingelser) givet ved

$$K = y^{\alpha} e^{-\beta y} x^{\gamma} e^{-\delta x}.$$

Antag, at $\alpha = 2/3$, $\beta = 4/3$, $\gamma = \delta = 1$, og at rov- og byttedyr bestandsstørrelserne er blevet observeret til henholdsvis $y = 1/2$ og $x = 1$. Usikkerhederne på observationerne antages at være $\sigma_y^2 = 1/16^2$ og $\sigma_x^2 = 1/8^2$, og endvidere antages observationerne uafhængige.

Spørgsmål VI.1 (15)

Hvad er approksimationen af variansen af K beregnet ved hjælp af den ikke-lineære fejlophobningsregel?

1 ☐ 0.312

2 ☐ 0

3 ☐ 0.559

4 ☐ 0.75

5 ☐ 0.889

Fortsæt på side 13

Spørgsmål VI.2 (16)

Antag nu, at rovdyrbestanden observeres uden fejl ($\sigma_y^2 = 0$), og dermed er den eneste kilde til usikkerhed byttedyrsbestanden. Ved brug af variansen ovenfor ($\sigma_x^2 = 1/8^2$) og under antagelse af normal fordelte fejl i x, dvs. $X = x+\epsilon$ med $\epsilon \sim N(0, \sigma_x^2)$, i hvilket interval falder standardafvigelsen for K så (svaret bør ikke baseres på de ikke-lineære tilnærmelser af fejlophobningsreglen, men bør baseres på simulering)?

1 ☐ $(0.07, 0.1)$

2 ☐ $(0.12, 0.2)$

3 ☐ $(10^{-4}, 0.01)$

4 ☐ $(0.03, 0.05)$

5 ☐ $(0.3, 0.5)$

Fortsæt på side 14

Opgave VII

Lad funktionen $f(x)$ være defineret ved

$$f(x) = \alpha\phi_1(x) + \beta\phi_2(x),$$

hvor $\phi_i(x)$ er tæthedsfunktionen for en normalfordelt stokastisk variabel med middelværdi μ_i og varians σ_i^2.

Spørgsmål VII.1 (17)

Under hvilke betingelser er $f(x)$ en tæthedsfunktion (svaret skal gælde for enhver værdi af $\sigma_i > 0$ og $\mu_i \in \mathbb{R}$)?

1 ☐ $\alpha = \beta = 1$

2 ☐ $\alpha \in [0, 2]$ og $\beta = 2 - \alpha$

3 ☐ $\alpha = \frac{\sigma_1^2}{\sigma_2^2}$ og $\beta = \frac{\sigma_2^2}{\sigma_1^2}$

4 ☐ $\alpha \in [0, 1]$ og $\beta = 1 - \alpha$

5 ☐ $\alpha = \frac{\mu_1^2}{\sigma_1^2}$ og $\beta = \frac{\mu_2^2}{\sigma_2^2}$

Fortsæt på side 15

Opgave VIII

En flyproducent bruger en dyr type skruer i produktionen af en bestemt model. For at reducere produktionsomkostningerne overvejer producenten at erstatte de dyre skruer med en billigere type skruer. Derfor tester producenten trækstyrken (MPa) af de to typer skruer, og resultaterne er vist i nedenstående tabel.

Trækstyrke	Billig	Dyr
Stikprøvegennemsnit (MPa)	1250	1300
Stikprøve standard afvigelse (MPa)	54.24	28.54
Stikprøvestørrelse	25000	15000

Spørgsmål VIII.1 (18)

Under antagelse af at prøverne var fuldstændig tilfældige, hvad er 95% konfidensintervallet for forskellen i middelværdien af trækstyrker (middelværdi for den billige type minus middelværdi for den dyre type) baseret på testresultaterne?

1 ☐ $[-50.07; -49.93]$

2 ☐ $[-50.13; -49.87]$

3 ☐ $[-50.34; -49.66]$

4 ☐ $[-50.68; -49.32]$

5 ☐ $[-50.81; -49.19]$

Spørgsmål VIII.2 (19)

Under nulhypotesen $H_0 : \mu_{\text{billig}} - \mu_{\text{dyr}} = -50$, hvad er den observerede teststørrelse baseret på testresultaterne?

1 ☐ -241.14

2 ☐ -120.57

3 ☐ -2.31

4 ☐ 0.00

5 ☐ 241.14

Producenten overvejer også at købe en ny og mere brændstoføkonomisk flymodel, og derfor har de målt de to modellers brændstofforbrug (i kg) på 10 populære ruter under lignende vejr- og vægtforhold. Producenten er kun interesseret i at sammenligne logaritmen af brændstofforbruget som angivet i nedenstående tabel:

Logaritmen af brændstofforbrug	Nuværende model	Ny model
Sapporo - Tokyo	7.964	7.932
Sydney - Melbourne	7.813	7.762
Mumbai - Delhi	8.299	8.243
Beijing - Shanghai	8.219	8.174
Paris - Montreal	9.832	9.782
Dubai - London	9.829	9.775
London - New York	9.842	9.794
New York - Los Angeles	9.498	9.445
Kuala Lumpur - Singapore	7.023	6.942
Cancun - Mexico City	8.408	8.347

Data kan læses i R ved hjælp af nedenstående kode (`c` for nuværende (current) og `n` for ny (new)):

```
log_c <- c(7.964,7.813,8.299,8.219,9.832,9.829,9.842,9.498,7.023,8.408)
log_n <- c(7.932,7.762,8.243,8.174,9.782,9.775,9.794,9.445,6.942,8.347)
```

Spørgsmål VIII.3 (20)

Hvad er p-værdien for den passende test af nulhypotesen $H_0 : \delta = \mu_{\text{nuværende}} - \mu_{\text{ny}} = 0.05$ mod en tosidet alternativ hypotese? (Her refererer $\mu_{\text{nuværende}}$ og μ_{ny} til middelværdien af logaritmen af brændstofforbruget.)

1 ☐ $p < 0.001$

2 ☐ $p = 0.442$

3 ☐ $p = 0.452$

4 ☐ $p = 0.908$

5 ☐ $p = 0.995$

Fortsæt på side 17

Opgave IX

En bilproducent vil gerne finde ud af, om der er forskel på brudstyrken i metalstænger lavet med metal fra forskellige leverandører. Lad Y repræsentere brudstyrken af metalstænger. I det følgende er brudstyrken målt på metalstænger hver fremstillet med metal fra en enkelt leverandør. Metal fra fire forskellige leverandører var inkluderet i undersøgelsen, og brudstyrken blev målt for 5 metalstænger fra hver leverandør:

Leverandør A	Leverandør B	Leverandør C	Leverandør D
92.0	131.0	74.1	90.4
111.6	103.5	52.8	95.2
98.4	100.0	82.5	87.6
87.7	84.7	94.7	63.2
134.9	134.5	107.3	119.5

Spørgsmål IX.1 (21)

Ingeniørerne i virksomheden har udført følgende analyse i R. Hvad er konklusionen, på signifikansniveau $\alpha = 5\%$, om forskel i brudstyrken af teststængerne lavet med metal fra de forskellige leverandører (både konklusion og argument skal være korrekt)?

```
anova(lm(y ~ Leverandør))

## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value Pr(>F)
## Leverandør  3 2508.8  836.25   2.027 0.1507
## Residuals  16 6601.0  412.56
```

1 ☐ En signifikant forskel i brudstyrken er ikke påvist, da p-værdien er større end signifikansniveauet.

2 ☐ En signifikant forskel i brudstyrken er påvist, da p-værdien er større end signifikansniveauet.

3 ☐ En signifikant forskel i brudstyrken er ikke påvist, da p-værdien er mindre end signifikansniveauet.

4 ☐ En signifikant forskel i brudstyrken er påvist, da p-værdien er mindre end signifikansniveauet.

5 ☐ Ingen af ovenstående konklusioner er korrekt.

Ingeniørerne huskede nu, at styrkeprøverne blev lavet på forskellige dage, og at forholdene (f.eks. vejret) kunne være forskellige mellem dagene. Heldigvis havde nogen allerede tænkt over

dette, og hver dag var netop en metalstang fra hver leverandør testet. Derfor kunne de gruppere observationerne på dage.

	Leverandør A	Leverandør B	Leverandør C	Leverandør D
Dag 1	92.0	131.0	74.1	90.4
Dag 2	111.6	103.5	52.8	95.2
Dag 3	98.4	100.0	82.5	87.6
Dag 4	87.7	84.7	94.7	63.2
Dag 5	134.9	134.5	107.3	119.5

Dette resulterede i følgende analyse (bemærk, nogle af værdierne i resultatet er erstattet af bogstaver og eventuelle $*$ i resultatet er blevet fjernet):

```
anova(lm(y ~ Leverandør + Dag))

## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Leverandør   3   A     836.25  3.8696    E
## Dag          4   B    1001.92  4.6362    F
## Residuals   12   C       D
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål IX.2 (22)

Hvilken konklusion kan drages, på signifikansniveauet $\alpha = 5\%$, fra denne analyse (både konklusion og argumentation skal være korrekt)?

1 ☐ Der er ikke en signifikant effekt af hverken leverandør eller dag, da de relevante p-værdier er henholdsvis 0.076 og 0.141.

2 ☐ Der er en signifikant effekt af leverandør, men ikke af dag, da de relevante p-værdier er henholdsvis 0.038 og 0.141.

3 ☐ Der er en signifikant effekt af både leverandør og dag, da de relevante p-værdier er henholdsvis 0.038 og 0.017.

4 ☐ Der er ikke en signifikant effekt af leverandør, men der er en signifikant effekt af dag, da de relevante p-værdier er henholdsvis 0.076 og 0.017.

5 ☐ Der er ikke en signifikant effekt af hverken leverandør eller dag, da de relevante p-værdier er henholdsvis 0.892 og 0.112.

Spørgsmål IX.3 (23)

Hvad er den totale kvadratafvigelsessum (SST)?

1 ☐ 216

2 ☐ 2509

3 ☐ 4008

4 ☐ 6516

5 ☐ 9110

Fortsæt på side 20

Opgave X

En person (investor 1) beslutter at investere i aktier. Investoren investerer 10000 kr. i 10 forskellige aktier (1000 kr. i hver), efter et år sælger investoren alle aktierne. Afkastet (målt i kr.) af de 10 aktier er angivet i nedenstående tabel.

Aktie	1	2	3	4	5	6	7	8	9	10
Afkast	1144	1218	1480	747	1178	-121	-382	-24	-652	-32

Investoren ønsker at undersøge, om investeringen har været en succes. Succeskriteriet er, at afkastet er signifikant forskellig fra (ved brug af en to-sidet test) og større end et 2% (200 kr. ialt, eller 20 kr. pr. aktie) afkast ved brug af signifikansniveau $\alpha = 5\%$.

Investoren har valgt at bruge en test uden nogen fordelingsantagelse for populationen, det antages dog at afkastene er uafhængige. Noget af R-koden nedenfor skal bruges til næste spørgsmål.

```
x <- c(1144,  1218,   1480,    747,  1178, -121, -382,  -24, -652, -32)

k <- 10000
sim.samp <- replicate(k, sample(x, replace = TRUE))

quantile(apply(sim.samp, 2, mean), prob = c(0.025, 0.05, 0.95, 0.975),
         type = 2)

##   2.5%     5%    95%  97.5%
##   0.15  68.60 832.95 905.40

t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 1.8531, df = 9, p-value = 0.09687
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -100.5569 1011.7569
## sample estimates:
## mean of x
##     455.6

t.test(x, mu = 200)

##
##  One Sample t-test
##
## data:  x
```

```
## t = 1.0396, df = 9, p-value = 0.3256
## alternative hypothesis: true mean is not equal to 200
## 95 percent confidence interval:
##  -100.5569 1011.7569
## sample estimates:
## mean of x
##     455.6
```

Spørgsmål X.1 (24)

Hvad er resultatet af testen (både konklusionen og argumentet skal være korrekt)?

1 ☐ Null-hypothesen accepteres, da p-værdien er 0.3

2 ☐ Null-hypothesen accepteres, da $20 \in [0.15, 905.4]$

3 ☐ Null-hypothesen forkastes, da p-værdien er 0.1

4 ☐ Null-hypothesen forkastes, da $0 < 0.15$

5 ☐ Null-hypothesen forkastes, da $0 \in [-101, 1012]$

En anden investor har investeret i 10 andre aktier, ligeledes med 1000 kr. i hvert aktie. De to investorer ønsker at sammenligne afkastet af deres investeringer. I R-koden nedenfor er `x` afkastet for investor 1 og `y` er afkastet for investor 2.

```
k <- 10000
sim.x <- replicate(k, sample(x, replace = TRUE))
sim.y <- replicate(k, sample(y, replace = TRUE))
sim.diff <- replicate(k, sample(x - y, replace = TRUE))

quantile(apply(sim.x, 2, mean) - apply(sim.y, 2, mean), prob = c(0.025,0.975),
         type = 2)

##        2.5%       97.5%
## -1562.82480    99.26984

quantile(apply(sim.diff, 2, mean), prob = c(0.025,0.975))

##       2.5%      97.5%
## -1186.2022  -285.1908

t.test(x, y, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = -3.0451, df = 9, p-value = 0.0139
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1298.7101  -191.5961
## sample estimates:
## mean difference
##       -745.1531
```

De to investorer ønsker at sammenligne afkastet af deres investeringer ved hjælp af en statistisk metode med så få antagelser som muligt, f.eks. uden nogen fordelingsantagelser overhovedet, det antages dog at alle observationer er uafhængige.

Spørgsmål X.2 (25)

Er der, baseret på R-output ovenfor, en signifikant (ved brug af signifikansniveau $\alpha = 5\%$) forskel mellem de to investores afkast (både konklusionen og argumentet skal være korrekt)?

1 ☐ Ja, da $0 \notin [-1186, -285]$

2 ☐ Ja, da $0.014 < 0.05$

3 ☐ Ja, da $0 \notin [-1299, -192]$

4 ☐ Nej, da $0.014 < 0.05$

5 ☐ Nej, da $0 \in [-1562, 99]$

Fortsæt på side 23

Opgave XI

Ingeniørerne hos et byggefirma er ved at udvikle en model for bygningers levetid. Modellen kan skrives som

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i,$$

hvor Y_i (`Life`) er levetiden (i år) for bygning i, x_{1i} (`Econ_RF`) den økonomiske risikofaktor for bygning i, x_{2i} (`Matr_RF`) er materiale risikofaktoren for bygning i, og x_{3i} (`Dsgn_RF`) designrisikofaktoren for bygning i. Endvidere antager modellen, at fejlene, ϵ_i, er uafhængige og identisk fordelt med en $N(0, \sigma^2)$-fordeling.

Ingeniørerne har estimeret modellen ved hjælp af mindste kvadraters metode, og output fra R er givet nedenfor:

```
##
## Call:
## lm(formula = Life ~ Econ_RF + Matr_RF + Dsgn_RF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5709  -3.3908  -0.0053   3.3407  15.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 151.76054    1.20254 126.199   <2e-16 ***
## Econ_RF      -0.35638    0.01270 -28.055   <2e-16 ***
## Matr_RF      -0.90717    0.01398 -64.878   <2e-16 ***
## Dsgn_RF      -0.11806    0.01289  -9.156   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.992 on 165 degrees of freedom
## Multiple R-squared:  0.9681,Adjusted R-squared:  0.9675
## F-statistic:  1669 on 3 and 165 DF,  p-value: < 2.2e-16
```

Du kan antage, at modellens forudsætninger er opfyldt.

Fortsæt på side 24

Spørgsmål XI.1 (26)

En bestemt bygning havde en levetid på 81.3 år. Bygningen havde en økonomisk risikofaktor på 64.7, en materiale risikofaktor på 55.2 og en designrisikofaktor på 28.1. Hvad er residualet for denne bygning?

1 ☐ Residualet kan ikke bestemmes uden yderligere oplysninger

2 ☐ 6.0 år

3 ☐ -6.0 år

4 ☐ 5.0 år

5 ☐ -5.0 år

Spørgsmål XI.2 (27)

Hvor mange bygninger (observationer) var inkluderet i datasættet, der blev brugt til at estimere modellen?

1 ☐ 165

2 ☐ 166

3 ☐ 167

4 ☐ 168

5 ☐ 169

Fortsæt på side 25

Opgave XII

En ung professionel skakspiller skal efter planen spille 20 partier den kommende måned. Partierne kan antages at være uafhængige, og spilleren antages at have samme sandsynlighed for at vinde hvert parti. Lad X angive antallet af partier, som spilleren vinder den kommende måned.

Spørgsmål XII.1 (28)

Hvad er den passende statistiske model/fordeling for X?

1 ☐ En binomialfordeling

2 ☐ En χ^2-fordeling

3 ☐ En F-fordeling

4 ☐ En hypergeometrisk fordeling

5 ☐ En Poissonfordeling

Spørgsmål XII.2 (29)

Den unge spiller har opnået følgende resultater i løbet af de sidste tre år:

År	Vundne	Uafgjorte	Tabte
2023	43	78	33
2022	25	55	22
2021	34	46	41

Spilleren ønsker nu at teste, om resultatfordelingen har ændret sig gennem årene, dvs. teste nulhypotesen om, at andelen af sejre, uafgjorte og tab er de samme over de tre år.

Når man udfører den sædvanlige hypotesetest af nulhypotesen, hvilken fordeling skal så bruges til at beregne den kritiske værdi?

1 ☐ En $\chi^2(4)$-fordeling

2 ☐ En $\chi^2(9)$-fordeling

3 ☐ En $\mathrm{F}(2,2)$-fordeling

4 ☐ En $\mathrm{F}(3,3)$-fordeling

5 ☐ En $\mathrm{F}(2,6)$-fordeling

Spørgsmål XII.3 (30)

Spilleren og en ven sammenligner deres præstationsvurderinger på tværs af syv turneringer, de begge deltog i. Ved hver turnering får de begge en præstationsvurdering baseret på deres resultater, og de antager, at deres præstationer ved forskellige turneringer er uafhængige. Men når de ser på deres præstationsvurderinger i nedenstående tabel, bemærker de, at deres præstationer ser ud til at afhænge af niveauet af turneringerne (lokal, national eller international).

Turnering	Lokal 1	Lokal 2	Lokal 3	Lokal 4	National 1	National 2	International 1
Spiller	2219	2248	2311	2256	2175	2140	2025
Ven	2144	2169	2341	2222	2088	2055	1979

Hvis man antager at observationerne er normalfordelte, hvilken af følgende statistiske test er så den mest passende til at sammenligne spillerens og vennens gennemsnitlige præstationer?

1 ☐ En Welch to-stikprøve t-test

2 ☐ En parret t-test

3 ☐ En sammenvægtet to-stikprøve t-test

4 ☐ En en-vejs ANOVA med to behandlinger (gruppe)

5 ☐ En χ^2-test

SÆTTET ER SLUT. God sommer!