Danmarks Tekniske Universitet *Side 1 af 24 sider.*

Skriftlig prøve: 24. June 2022

Kursus navn og nr.: **Introduktion til Matematisk Statistik (02403)**

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 8 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" svararket (6 separate sider) på Digital Eksamen med numrene på de svarmuligheder, som du mener er de rigtige.

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 svararket online via CampusNet. 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	I.4	I.5	II.1	II.2	II.3	II.4	III.1
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

Opgave	VI.2	VI.3	VI.4	VI.5	VI.6	VII.1	VIII.1	VIII.2	VIII.3	VIII.4
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Eksamenssættet består af 24 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

Fødselsvægten af 50 nyfødte piger er blevet rapporteret i et ukendt land, og gennemsnittet samt spredningen er beregnet til hhv. $\bar{x}_{\text{p}} = 3505.7$ g and $s_{\text{p}} = 467.9$ g.

Spørgsmål I.1 (1)

Hvad er 95%-konfidensintervallet for middelværdien af fødselsvægten af piger, μ_{p}?

1 ☐ $[3328.3, 3683.0]$

2 ☐ $[3372.7, 3638.7]$

3 ☐ $[3371.4, 3640.0]$

4 ☐ $[3328.4, 3683.0]$

5 ☐ $[3499.6, 3511.8]$

Spørgsmål I.2 (2)

Den gennemsnitlige fødselsvægt af danske piger er opgivet til 3449 g. Vi vil teste om fødselsvægten i det ukendte land afviger signifikant fra fødselsvægten af danske piger, så vi tester hypotesen $H_0 : \mu_{\text{p}} = 3449$ g ved brug af det observerede data.

Den tilhørende teststatistik er beregnet til $t_{\text{obs}} = 0.857$. Hvilke af de følgende udsagn er korrekte, når vi benytter et signifikansniveau på $\alpha = 0.05$ (både p-værdi og konklusion skal være korrekte)?

1 ☐ p-værdi $= 0.198$, og hypotesen kan ikke afvises.

2 ☐ p-værdi $= 0.198$, og hypotesen afvises.

3 ☐ p-værdi $= 0.396$, og hypotesen kan ikke afvises.

4 ☐ p-værdi $= 1.604$, og hypotesen afvises.

5 ☐ p-værdi $= 0.396$, og hypotesen afvises.

Fortsæt på side 3

Fødselsvægten af 50 nyfødte drenge er også blevet rapporteret (i det samme ukendte land). Gennemsnittet og spredningen er beregnet til $\bar{x}_{\mathrm{d}} = 3619.4$ g og $s_{\mathrm{d}} = 409.0$ g. Vi ønsker at teste hypotesen om at middelværdierne af fødselsvægten af piger og drenge er ens mod den alternative hypotese om at middelværdierne er forskellige. Der bruges et signifikansniveau på $\alpha = 0.05$ i de resterende spørgsmål i opgaven.

Spørgsmål I.3 (3)

Under nulhypotesen om, at der ikke er nogen forskel på middelværdierne af fødselsvægten af drenge og piger, følger (Welch) teststatistikken for 2 stikprøver en t-fordeling med ν frihedsgrader. Hvad er ν i vores tilfælde lig med?

1 ☐ 100

2 ☐ 98.24

3 ☐ 49

4 ☐ 98

5 ☐ 96.28

Fortsæt på side 4

Spørgsmål I.4 (4)

Antag at antallet af frihedsgrader, ν, er gemt i R i variablen `v`. Hvilken kommando resulterer i den korrekte kritiske værdi i t-fordelingen, der skal benyttes i hypotesetesten fra forrige spørgsmål?

1 ☐
```
qt(0.975,v)
```

2 ☐
```
1-pt(0.975,v)
```

3 ☐
```
pt(0.95,v)
```

4 ☐
```
1-qt(0.95,v)
```

5 ☐
```
qt(0.95,v)
```

Fortsæt på side 5

Spørgsmål I.5 (5)

De rapporterede fødselsvægte af piger og drenge er i R gemt i hhv. `xp` og `xd`. Hvilken af de nedenstående kommandoer genererer det korrekte konfidensinterval for forskellen på middelværdierne?

1 ☐
```r
t.test(xp, xd, paired = TRUE)
```

2 ☐
```r
t.test(xp, xd, paired = TRUE, conf.level = 0.90)
```

3 ☐
```r
t.test(xp, xd, conf.level = 0.90)
```

4 ☐
```r
t.test(xp, xd, paired = TRUE, conf.level = 0.95)
```

5 ☐
```r
t.test(xp, xd)
```

Fortsæt på side 6

Opgave II

En dansk virksomhed ønsker at undersøge, om medarbejdernes faglige kompetencer på en virtual reality-platform (VR) påvirker deres opgavekvalitet. 200 medarbejdere deltog. Følgende optællingsdata giver et overblik over opgavekvaliteten (dårlig, middel og god) versus niveauet for VR-træningsengagement (under gennemsnit, gennemsnit og over gennemsnit).

Opgavekvalitet / VR-træningsengagement	Under gennemsnit	Gennemsnit	Over gennemsnit	Række total
Dårlig	11	27	15	53
Middel	14	40	30	84
God	5	23	35	63
Kolonne total	30	90	80	200

Nulhypotesen om uafhængighed mellem opgavekvalitet og VR-træningsengagement skal testes med χ^2-test.

Medarbejdere med "middel" og "god" opgavekvalitet betragtes som "effektive medarbejdere".

Spørgsmål II.1 (6)

Hvad er det forventede antal personer med "under gennemsnit" VR-træningsengagement og "dårlig opgavekvalitet" under H_0 (dvs. vi antager H_0 er sand)?

1 ☐ 7.95

2 ☐ 21.83

3 ☐ 25.2

4 ☐ 19.43

5 ☐ 9.45

Fortsæt på side 7

Spørgsmål II.2 (7)

Hvad er 95% konfidensintervallet for andelen af "Effektive medarbejdere" baseret på data givet ovenfor?

1 ☐ $[0.674, 0.796]$

2 ☐ $[0.621, 0.749]$

3 ☐ $[0.532, 0.668]$

4 ☐ $[0.426, 0.578]$

5 ☐ $[0.706, 0.824]$

Spørgsmål II.3 (8)

Hvad er 95% konfidensintervallet for forskellen i andelen af medarbejdere med "god" opgavekvalitet med VR-træningsengagement "over gennemsnit" og "gennemsnit" ($p_{\text{Over Gns, god}} - p_{\text{Gns, god}}$)?

1 ☐ $[0.019, 0.361]$

2 ☐ $[0.043, 0.212]$

3 ☐ $[0.011, 0.313]$

4 ☐ $[0.041, 0.323]$

5 ☐ $[0.044, 0.091]$

Som en hjælp til det næste spørgsmål er følgende R-kode, hvor `training` er antalstabellen, blevet kørt (nogle tal er blevet erstattet af bogstaver):

```
chisq.test(training, correct = FALSE)
##
## Pearson's Chi-squared test
##
##  data:  training
## X-squared = 10.985, df = A, p-value = B
```

Fortsæt på side 8

Spørgsmål II.4 (9)

Betragter vi χ^2-teststørrelsen, hvad bliver da p-værdien og den korrekte konklusion ved signifikansniveau $\alpha = 0.05$ (alle dele af svaret skal være korrekte)?

1 ☐ Der er en signifikant sammenhæng mellem VR-træningsengagement og opgavekvalitet, da p-værdien $= 0.027 < 0.05 = \alpha$

2 ☐ Der er ikke påvist en signifikant sammenhæng mellem VR-træningsengagement og opgavekvalitet, da p-værdien $= 0.027 < 0.05 = \alpha$

3 ☐ Nul-hypotesen kan ikke forkastes, da p-værdien $= 0.50 > \alpha = 0.05$

4 ☐ Der er en signifikant sammenhæng mellem VR-træningsengagement og opgavekvalitet, da p-værdien $= 0.037 < 0.05 = \alpha$

5 ☐ Der er ikke påvist en signifikant sammenhæng mellem VR-træningsengagement og opgavekvalitet, da p-værdien $= 0.037 < 0.05 = \alpha$.

Fortsæt på side 9

Opgave III

I et studie om forgiftning af rotter, målte man overlevelsestiden (i dage) for 24 rotter. Hver rotte fik gift og blev efterfølgende behandlet med en af fire behandlinger, A, B, C og D. Lad Y betegne logaritmen til overlevelsestiden (`logt`), som bruges i analysen.

Man har anvendt en envejs ANOVA på data:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \text{ hvor } \varepsilon_{ij} \sim N(0, \sigma^2) \text{ og i.i.d.}$$

```
  logt <- c(-1.02, -1.24, -0.92, -1.47, -0.08, -0.49, -0.71, 0.22, -0.82, -1.05,
            -1.17, -0.92, -0.58,  0.02, -0.34, -0.97, -1.51, -1.56, -1.2, -0.99,
            -1.47, -1.39, -1.2,  -1.02)
treatments <-
  as.factor(c("A", "A", "A", "A", "B", "B", "B", "B", "C", "C", "C", "C",
              "D", "D", "D", "D", "A", "A", "B", "B", "C", "C", "D", "D"))
```

Desuden oplyses det at $SS(Tr) = 2.286$ og $SSE = 3.241$, og at gruppegennemsnittene er:

```
tapply(logt, treatments, mean)

##          A          B          C          D
## -1.2866667 -0.5416667 -1.1366667 -0.6816667
```

Spørgsmål III.1 (10)

Hvad er estimatet for effekten af behandling B, $\hat{\alpha}_B$?

1 ☐ -0.542

2 ☐ 0.370

3 ☐ 0.542

4 ☐ 2.22

5 ☐ 2.33

Fortsæt på side 10

Spørgsmål III.2 (11)

Hvad er værdien af den sædvanlige teststørrelse (F), for test af forskel i behandlinger?

1 ☐ 0.0121

2 ☐ 0.7051

3 ☐ 4.702

4 ☐ 14.11

5 ☐ 16.93

Forskerne er særligt interesserede i at sammenligne behandling B og D, da de med deres viden indenfor kemi forventer at disse behandlinger bør være omtrent lige gode.

Spørgsmål III.3 (12)

Hvad er konklusionen om post hoc-forskellen i middelværdi mellem behandling B og D på et 5% signifikansniveau (både konklusion og argument skal være korrekt)?

1 ☐ 95% konfidensintervallet for forskellen i middelværdi er $[-0.902, 0.622]$. Der er dermed ikke signifikant forskel mellem behandlingerne.

2 ☐ 95% konfidensintervallet for forskellen i middelværdi er $[-0.659, 0.338]$. Der er dermed ikke signifikant forskel mellem behandlingerne.

3 ☐ 95% konfidensintervallet for forskellen i middelværdi er $[-0.659, 0.338]$. Der er dermed signifikant forskel mellem behandlingerne.

4 ☐ 95% konfidensintervallet for forskellen i middelværdi er $[-0.625, 0.345]$. Der er dermed ikke signifikant forskel mellem behandlingerne.

5 ☐ 95% konfidensintervallet for forskellen i middelværdi er $[-0.206, -0.020]$. Der er dermed signifikant forskel mellem behandlingerne.

Fortsæt på side 11

Opgave IV

I en kontorbygning blev den tid hvor et rum var tilgængeligt (dvs. tomt) målt gennem en periode på ca. 8 måneder, der er en observation hver gang rummet skifter fra optaget til tomt (det betyder, at der kan være mere end en observation pr. dag). Tilgængelighedsvarighed måltes som ledige timer indenfor normal kontortid, og målingerne blev gemt i vektoren `time`. Et sammendrag ("summary") af målte tider for tilgængelighed er givet nedenfor.

```
summary(time)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2500  0.9375  2.5000  2.7017  3.7500 11.7500
```

Spørgsmål IV.1 (13)

Hvad er kvartilbredden (Inter Quartile Range, IQR) for de præsenterede data?

1 ☐ 0.20

2 ☐ 2.70

3 ☐ 2.50

4 ☐ 2.81

5 ☐ 11.5

Antag, at en stokastisk variabel X følger en eksponential fordeling med middelværdi lig med det observerede gennemsnit af tiderne for tilgængelighed.

Spørgsmål IV.2 (14)

Hvad er medianen af X?

1 ☐ 2.50

2 ☐ 0.169

3 ☐ 1.87

4 ☐ 2.70

5 ☐ 0.741

Fortsæt på side 12

Spørgsmål IV.3 (15)

Stadig ander antagelse af den eksponentielle fordeling med middelværdi lig med den observerede gennemsnitlige tilgængelighedstid, hvilket af følgende stykker R-kode beregner et 95% parametrisk bootstrap-konfidensinterval for den forventede værdi af tilgængelighedstiden (i alle tilfælde er `n = length(tid)` og `k`=10^4)?

1 ☐
```
m <- mean(time)
X <- matrix(rexp(n * k, m), ncol=k)
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

2 ☐
```
m <- mean(time)
X <- matrix(rexp(n * k, 1/m), ncol=k)
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

3 ☐
```
X <- replicate(k, sample(time, replace = TRUE))
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

4 ☐
```
X <- replicate(n, sample(time, replace = TRUE, size = k))
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

5 ☐
```
m <- mean(time)
X <- matrix(rnorm(n * k, m, sd(time)), n)
quantile(apply(X, 2, mean), prob = c(0.025, 0.975))
```

Fortsæt på side 13

Det er af interesse at undersøge, om variationskoefficienten er lig med 1. Til det formål er følgende R-kode blevet evalueret (inklusive resultaterne):

```
k <- 1e4
n <- length(time)

X <- replicate(k, sample(time, replace = TRUE))
quantile(apply(X, 2, sd) / apply(X, 2, mean),
         prob = c(0.025, 0.05, 0.95, 0.975))

##      2.5%        5%       95%     97.5%
## 0.7586455 0.7727351 0.9228017 0.9382224

quantile(apply(X, 2, var) / apply(X, 2, mean),
         prob = c(0.025, 0.05, 0.95, 0.975))

##     2.5%       5%      95%    97.5%
## 1.508020 1.570758 2.324900 2.400079

X2 <- replicate(k, rexp(n, m))
quantile( apply(X2, 2, sd) / apply(X2, 2, mean) ,
         prob = c(0.025, 0.05, 0.95, 0.975))

##      2.5%        5%       95%     97.5%
## 0.8767490 0.8933601 1.1094034 1.1368452

quantile(apply(X2, 2, var) / apply(X2, 2, mean),
         prob = c(0.025, 0.05, 0.95, 0.975))

##      2.5%        5%       95%     97.5%
## 0.2740744 0.2873999 0.4669260 0.4916159
```

Spørgsmål IV.4 (16)

Hvad kan vi, baseret på R-koden ovenfor, konkludere når signifikansniveau $\alpha = 0.05$ benyttes, og der ikke bruges nogen fordelingsantagelse (både konklusion og argument skal være korrekte)?

1 ☐ Det kan ikke afvises at variationskoefficienten er lig med 1, da $1 > 0.94$

2 ☐ Variationskoefficienten er mindre end 0.7 da et 95% konfidensinterval er $[0.27, 0.49]$

3 ☐ Det kan ikke afvises at variationskoefficienten er lig med 1, da $1 \notin [1.51, 2.4]$

4 ☐ Det kan ikke afvises at variationskoefficienten er lig med 1, da $1 \in [0.88, 1.14]$

5 ☐ Variationskoefficienten er ikke lig med 1, da $1 \notin [0.76, 0.94]$

Fortsæt på side 14

Et lignende sæt målinger blev taget i et andet rum, man ønsker at sammenligne middeltilgængelighedstiden mellem de to rum. Opsummeringen af data fra det andet rum er givet nedenfor.

```
summary(time2)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.250   1.500   2.500   2.551   3.250  14.250
```

Under antagelse af uafhængighed mellem rummene, blev det besluttet at teste, om der er en signifikant forskel i de forventede tilgængelighedstider mellem rummene ved brug af en test der ikke anvender nogen fordelingsantagelser.

Spørgsmål IV.5 (17)

Hvilket af følgende stykker R-kode kan benyttes til at teste hypotesen om, at der ikke er nogen forskel mellem den gennemsnitlige tilgængelighedstid for de to rum.

1 ☐
```
t.test(time, time2, paired = TRUE)
```

2 ☐
```
prop.test(sim1,sim2)
```

3 ☐
```
sim <- replicate(k, sample(time - time2, replace = TRUE))
quantile(apply(sim1, 2, mean), prob = c(0.025, 0.975))
```

4 ☐
```
sim1 <- replicate(k, sample(time, replace = TRUE))
sim2 <- replicate(k, sample(time2, replace = TRUE))
quantile(apply(sim1,2,mean) - apply(sim2,2,mean), prob = c(0.025,0.975))
```

5 ☐
```
t.test(time, time2)
```

Fortsæt på side 15

Opgave V

Produktionschefen på en papirfabrik har analyseret papir, der er blevet fremstillet af fabrikkens maskineri på forskellige dage. Han ønsker at teste om kvaliteten varierer som følge af maskine og produktionsdag.

Hans analyse gav følgende tovejs ANOVA-tabel (hvor nogle værdier er blevet udeladt):

```
## Analysis of Variance Table
##
## Response: quality
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## day        6 0.24902 0.041504  2.2731    X
## machine    3 0.21025 0.070085  3.8384    X
## Residuals 18 0.32866 0.018259
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Spørgsmål V.1 (18)

Hvor mange observationer er inkluderet i analysen?

1 ☐ 6

2 ☐ 9

3 ☐ 18

4 ☐ 27

5 ☐ 28

Fortsæt på side 16

Spørgsmål V.2 (19)

Hvad er konklusionen om effekt af produktionsdag på et 5% signifikansniveau (både konklusion og argument skal være korrekt)?

1 ☐ Der er en signifikant effekt, da p-værdien er 0.028

2 ☐ Der er ikke en signifikant effekt, da p-værdien er 0.083

3 ☐ Der er ikke en signifikant effekt, da p-værdien er 0.157

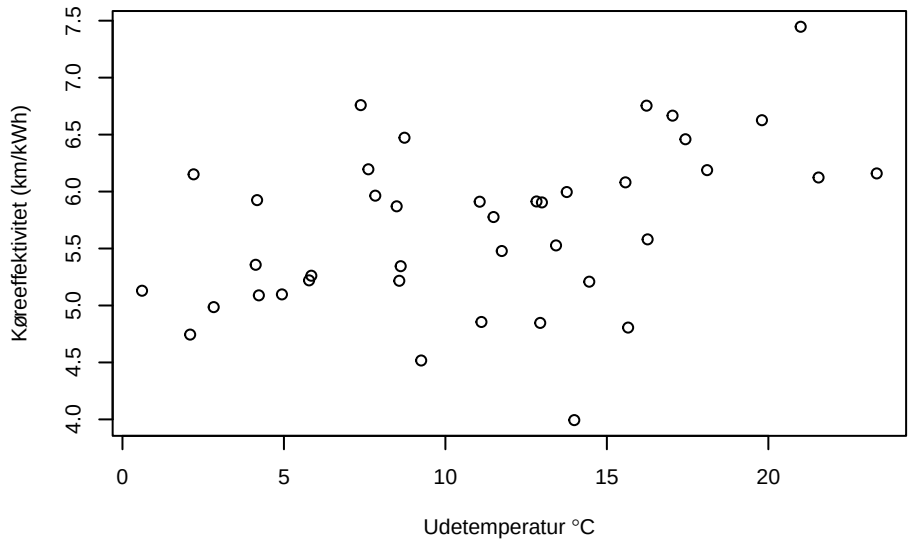
4 ☐ Der er ikke en signifikant effekt, da $SSE = 0.329$ er mellem ± 1.96, hvor 1.96 er 97.5%-fraktilen i en standard-normalfordeling, $N(0, 1)$.

5 ☐ Vi har ikke tilstrækkelig information til at konkludere hvorvidt produktionsdag er signifikant eller ej.

Fortsæt på side 17

Opgave VI

Ejeren af en elbil ønskede at finde ud af hvilken effekt udetemperaturen har på bilens rækkevidde. Så hun målte køreeffektiviten (rækkevidde per energienhed) samt udetemperaturen, på hver tur hun foretog i en periode. Data kan ses i nedenstående scatter-plot:



En simpel lineær regressionsmodel med køreeffektiviteten som den afhængige variable, og udetemperaturen som den forklarende variabel, blev anvendt. Resultaterne var:

```
##
## Call:
## lm(formula = Effektivitet ~ Udetemperatur)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84694 -0.27181  0.01402  0.43993  1.26562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.10493    0.22551  22.638   <2e-16 ***
## Udetemperatur  0.05259    0.01799   2.924   0.0058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6573 on 38 degrees of freedom
## Multiple R-squared:  0.1836,Adjusted R-squared:  0.1622
## F-statistic: 8.548 on 1 and 38 DF,  p-value: 0.0058
```

Spørgsmål VI.1 (20)

Hvilket af følgende udsagn er korrekt (både konklusionen og argumentet skal være korrekt)?

1 ☐ På signifikansniveau $\alpha = 0.01$ kunne en signifikant korrelation mellem køreeffektiviteten og udetemperaturen påvises, da $0.0058 < 0.01$.

2 ☐ På signifikansniveau $\alpha = 0.01$ kunne en signifikant korrelation mellem køreeffektiviteten og udetemperaturen ikke påvises, da $0.05259 > 0.01$.

3 ☐ På signifikansniveau $\alpha = 0.05$ kunne en signifikant korrelation mellem køreeffektiviteten og udetemperaturen påvises, da $0.01799 < 0.05$.

4 ☐ På signifikansniveau $\alpha = 0.05$ kunne en signifikant korrelation mellem køreeffektiviteten og udetemperaturen ikke påvises, da $0.01799 < 0.05$.

5 ☐ På signifikansniveau $\alpha = 0.05$ kunne en signifikant korrelation mellem køreeffektiviteten og udetemperaturen ikke påvises, da $0.6573 > 0.05$.

Spørgsmål VI.2 (21)

Batteristørrelsen var på 54 kWh. Hvad er den prædikterede gennemsnitlige rækkevidde ved en udetemperatur på 5 °C ifølge modellen og de estimerede parametre?

1 ☐ 250 km

2 ☐ 260 km

3 ☐ 270 km

4 ☐ 280 km

5 ☐ 290 km

Fortsæt på side 19

Spørgsmål VI.3 (22)

Datapunktet $i = 5$ havde observation af udetemperaturen på 2.096 °C og en køreeffektivitet på 4.744 km/kWh. Hvad er residualet (dvs. den realiserede afvigelse) for dette datapunkt?

1 ☐ -0.471

2 ☐ 0.226

3 ☐ 0.657

4 ☐ 0.843

5 ☐ 1.634

Spørgsmål VI.4 (23)

Bilejeren ønskede at undersøge luftfugtighedens effekt på rækkevidden. Hun fik derfor fat i observationer af luftfugtigheden fra en nærliggende vejrstation og matchede dem med hendes observationer.

Hun ønskede at anvende en multipel lineær regressionsmodel med både udetemperaturen og luftfugtigheden som input, men før dette gjorde hun sig nogle overvejelser. Hvilket af følgende udsagn om anvendelsen af en multipel lineær regressionsmodel er ikke korrekt?

1 ☐ Det er oftest en god idé at undersøge scatterplots af alle mulige par af variablerne (et pairs plot i R).

2 ☐ Det er vigtigt af udføre en modelselektion.

3 ☐ Niveauet af korrelation mellem de forklarende variable kan ikke påvirke resultaterne.

4 ☐ Antallet af observationer påvirker resultaterne.

5 ☐ Det er vigtigt at udføre en modelvalidering med den udvalgte model.

Fortsæt på side 20

Spørgsmål VI.5 (24)

Hun anvendte en multipel lineær regressionsmodel med både udetemperatur og luftfugtighed. Resultatet var:

```
##
## Call:
## lm(formula = Effektivitet ~ Udetemperatur + Luftfugtighed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76776 -0.34382 -0.01327  0.38670  1.34920
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.96398    0.24226  20.490  < 2e-16 ***
## Udetemperatur  0.06454    0.01952   3.306  0.00211 **
## Luftfugtighed -0.16622    0.11379  -1.461  0.15250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 37 degrees of freedom
## Multiple R-squared:  0.2282,Adjusted R-squared:  0.1864
## F-statistic: 5.469 on 2 and 37 DF,  p-value: 0.008302
```

Hvad er konklusionen på et 'backward selection step' på 5% signifikansniveau for den anvendte multipel lineære regressionsmodel med dette resultat (både konklusion og argumentet skal være korrekt)?

1 ☐ Ingen af de to forklarende variable skal fjernes fra modellen, da $0.2282 > 0.05$.

2 ☐ Udetemperaturen skal fjernes fra modellen, da $0.00211 < 0.05$.

3 ☐ Udetemperaturen skal fjernes fra modellen, da $0.06454 > 0.05$.

4 ☐ Luftfugtigheden skal fjernes fra modellen, da $0.15250 > 0.05$.

5 ☐ Luftfugtigheden skal fjernes fra modellen, da $0.16622 > 0.05$.

Fortsæt på side 21

Modellen kan formuleres ved

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

hvor kolonnerne i $\boldsymbol{X}$ er: en vektor af et-taller, den målte udetemperatur, og den målte luftfugtighed.

Spørgsmål VI.6 (25)

Hvad er diagonal elementerne i matricen $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$?

1 ☐ $(0.374, 0.0301, 0.176)$

2 ☐ $(0.0906, 0.000588, 0.0200)$

3 ☐ $(0.612, 0.174, 0.419)$

4 ☐ $(0.140, 0.000908, 0.0309)$

5 ☐ $(0.577, 0.0465, 0.271)$

Fortsæt på side 22

Opgave VII

En familie spiller et gammelt spil kaldet 'mus'. I dette spil lægges 10 stykker slik på en tallerken. Et familiemedlem udvælges som spilleren og kigger væk, mens de andre peger på 2 stykker slik, som så kaldes musene. Spilleren vælger nu et stykke slik ad gangen. Hvis det valgte stykke er en mus, er turen forbi og spilleren beholder alle stykkerne samlet op indtil da. Det kan antages, at spilleren vælger stykkerne helt tilfældigt.

Spørgsmål VII.1 (26)

Hvad er sandsynligheden for at spilleren får alle 8 mulige stykker?

1 ☐ 2.2%

2 ☐ 3.6%

3 ☐ 5.8%

4 ☐ 6.4%

5 ☐ 9.2%

Fortsæt på side 23

Opgave VIII

Lad $X_i \sim LN(\mu, \sigma^2)$, $i = \{1, ..., n\}$ (dvs. $\log(X_i) \sim N(\mu, \sigma^2)$) være uafhængige stokastiske variable.

Spørgsmål VIII.1 (27)

Hvad er sandsynligheden $P(X_1 X_2 > k)$?

1 ☐ $1 - F(k)$, hvor F er fordelingsfunktionen for en log-normal fordeling med middelværdi μ og varians σ^2

2 ☐ $1 - F\left(\frac{k-2\mu}{2\sigma^2}\right)$, hvor F er fordelingsfunktionen for en standard normal fordelt stokastisk variabel

3 ☐ $1 - F(k)$, hvor F er fordelingsfunktionen for en log-normal fordeling med middelværdi 2μ og varians σ

4 ☐ $1 - F(k)$, hvor F er fordelingsfunktionen for en log-normal fordeling med middelværdi 2μ and standardafvigelse 2σ

5 ☐ $1 - F\left(\frac{\log k - 2\mu}{\sqrt{2}\sigma}\right)$, hvor F er fordelingsfunktionen for en standard normal fordelt stokastisk variabel.

Spørgsmål VIII.2 (28)

Størrelsen

$$Q = \left(\prod_{i=1}^{n} X_i\right)^{1/n} \tag{1}$$

kaldes også den geometriske middelværdi. Hvad er middelværdi og varians af Q?

1 ☐ $E[Q] = e^{\mu + \frac{1}{2}\frac{\sigma^2}{n}}$ og $V[Q] = \left(e^{\sigma^2/n} - 1\right) e^{2\mu + \frac{\sigma^2}{n}}$

2 ☐ $E[Q] = \mu$ og $V[Q] = \sigma^2/n$

3 ☐ $E[Q] = e^{\mu + \frac{1}{2}\sigma^2}$ og $V[Q] = \left(e^{\sigma^2} - 1\right) e^{2\mu + \sigma^2}$

4 ☐ $E[Q] = e^{\mu/n}$ og $V[Q] = e^{\sigma^2/n}$

5 ☐ $E[Q] = (\mu)^{1/n}$ og $V[Q] = (\sigma^2/n)^{1/n}$

Fortsæt på side 24

Spørgsmål VIII.3 (29)

Lad $X_i \sim N(\mu_1, \sigma_1^2)$ og $Y_j \sim N(\mu_2, \sigma_2^2)$ ($i = \{1, ..., n_1\}$ og $j = \{1, ..., n_2\}$) være uafhængige stokastiske variable. Lad desuden

$$Z_1 = \sum_{i=1}^{n_1} (X_i - \mu_1)^2$$
$$Z_2 = \sum_{j=1}^{n_2} (Y_j - \mu_2)^2$$

for hvilken værdi af σ_1^2 gælder der at $E[Z_1/Z_2] = 1$?

1 ☐ $\sigma_1^2 = \frac{\sigma_2^2 n_1 (n_2 - 2)}{n_2^2}$

2 ☐ $\sigma_1^2 = \frac{\sigma_2^2 (n_2 - 2)}{n_1}$

3 ☐ $\sigma_1^2 = \frac{\sigma_2^2 \mu_1 (n_1 - 1)(n_2 - 3)}{\mu_2 (n_2 - 1)^2}$

4 ☐ $\sigma_1^2 = \frac{\sigma_2^2 n_2}{n_1}$

5 ☐ $\sigma_1^2 = \frac{\sigma_2^2 \mu_1 (n_2 - 3)}{\mu_2 (n_2 - 1)}$

Spørgsmål VIII.4 (30)

Lad $X_i \sim N(0, 1)$ ($i = \{1, ..., n\}$) være "iid" stokastiske variable, og lad S^2 og $\bar{X}$ være defineret på den sædvanlige måde, hvad er fordelingen af $Q = n\bar{X}^2 + (n-1)S^2$?

1 ☐ $Q \sim F(1, n)$

2 ☐ $Q \sim N(0, n - 1)$

3 ☐ $Q \sim \chi^2(n - 1)$

4 ☐ $Q \sim F(1, n - 1)$

5 ☐ $Q \sim \chi^2(n)$

SÆTTET ER SLUT. God sommer!