Danmarks Tekniske Universitet *Side 1 af 25 sider.*

Skriftlig prøve: 16. December 2023

Kursus navn og nr.: **Introduktion til Statistik (02323)**

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

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

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

Eksamenssættet består af 25 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

En population er eksponentialfordelt med rate $\lambda = 2$.

Spørgsmål I.1 (1)

Hvilket af følgende udsagn er korrekt?

1 ☐ Sandsynligheden for at få en observation mellem 1 og 2 i en tilfældig trækning kan beregnes i R ved: `pexp(1, rate=2) - pexp(2, rate=2)`

2 ☐ Sandsynligheden for at få en observation mindre end 1 i en tilfældig trækning kan beregnes i R ved `pexp(2, rate=2)`

3 ☐ Sandsynligheden for at få en observation mindre end 1 i en tilfældig trækning kan beregnes i R ved: `1 - pexp(1, rate=2)`

4 ☐ Sandsynligheden for at få en observation større end 3 i en tilfældig trækning kan beregnes i R ved: `dexp(3, rate=2)`

5 ☐ Ingen af ovenstående udsagn er korrekte

Spørgsmål I.2 (2)

Ifølge den centrale grænseværdisætning (CLT) kan gennemsnittet af en tilfældig stikprøve med $n = 100$ observationer fra populationen tilnærmes med hvilken type fordeling (bemærk, at CLT ikke siger noget særligt om stikprøvestørrelsen $n = 100$)?

1 ☐ Standard normalfordelingen

2 ☐ En normalfordeling (som ikke også er en standard normalfordeling)

3 ☐ En eksponentialfordeling

4 ☐ En Poisson-fordeling

5 ☐ En F-fordeling

Fortsæt på side 3

Opgave II

Data fra et balanceret forsøg (lige mange observationer for hver behandling) modelleres med en ensidet variansanalyse. Resultaterne fra analysen kan findes i tabellen nedenfor, hvor nogle tal er blevet erstattet af bogstaver.

Kilde	Frihedsgrader	SS	MS	Teststørrelse	p-værdi
Behandling	9	207	D	E	0.03
Residualer	50	B	C		
Total	A	707			

Spørgsmål II.1 (3)

Hvilke værdier passer med tabellen?

1 ☐ $A = 59$, $B = 914$ og $D = 23$

2 ☐ $A = 59$, $C = 10$ og $E = 2.3$

3 ☐ $A = 450$, $D = 23$ og $E = 2.3$

4 ☐ $B = 500$, $C = 23$ og $D = 10$

5 ☐ $B = 914$, $C = 10$ og $E = 23$

Spørgsmål II.2 (4)

To specifikke behandlinger skal efterfølgende sammenlignes i en posthoc-analyse. Hvad er den mindste signifikante forskel (LSD) mellem behandlingernes gennemsnit på et 5% signifikansniveau?

1 ☐ 2.841

2 ☐ 3.060

3 ☐ 3.199

4 ☐ 3.667

5 ☐ 4.130

Fortsæt på side 4

Opgave III

Temperaturen indendørs er en vigtig del af indeklimaet og folks velbefindende, og desuden udgør opvarmning en væsentlig del af energiforbruget i huse.

En husejer betragter indendørstemperaturen i et værelse i sit hus. Til at starte med beslutter han at analysere den daglige gennemsnitstemperatur i værelset gennem nogen tid. R-outputtet fra hans analyse er givet nedenfor (vektoren `temp` indeholder de daglige gennemsnitstemperaturer i værelset).

```
##
##  One Sample t-test
##
## data:  temp
## t = 160.53, df = 233, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.97593 20.47234
## sample estimates:
## mean of x
##  20.22413
```

Spørgsmål III.1 (5)

Hvor mange dage har husejeren brugt i sin analyse?

1 ☐ 366

2 ☐ 364

3 ☐ 234

4 ☐ 365

5 ☐ 233

Spørgsmål III.2 (6)

Husejeren ønsker at teste en hypotese om, at gennemsnitstemperaturen i værelset er 20 °C mod alternativet, at gennemsnitstemperaturen er forskellig fra 20 °C. Hvad er den sædvanlige p-værdi for dennee hypotesetest?

1 ☐ $< 2.2 \cdot 10^{-16}$

2 ☐ 0.375

3 ☐ 0.0382

4 ☐ 0.137

5 ☐ 0.0765

Husejeren vil også gerne analysere variationen hen over tid. For at gøre dette beslutter han sig for at teste, hvorvidt gennemsnitstemperaturen på et givent tidspunkt på dagen er konstant over tid. Formelt set gør han dette ved at teste hypotesen om at temperaturen på det givne tidspunkt på dagen kan antages at være den samme i to forskellige måneder. Outputtet af denne analyse er (bemærk at teststørrelsen er blevet erstattet af `Q`):

```
## Welch Two Sample t-test
##
## data:  temp1 and temp2
## t = Q, df = 53.627, p-value = 0.9793
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9278637  0.9040722
## sample estimates:
## mean of x mean of y
##  19.10497  19.11686
```

Her er `temp1` og `temp2` vektorer med temperaturerne i de to forskellige måneder.

Spørgsmål III.3 (7)

Er der en signifikant forskel på gennemsnitstemperaturerne i de to måneder på signifikansniveauet $\alpha = 0.05$?

1 ☐ Ja, da $0.979 > 0.95$

2 ☐ Ja, da $0 \notin [19.10, 19.11]$

3 ☐ Nej, da $0.904 > 0.05$

4 ☐ Nej, da $0.979 > 0.05$

5 ☐ Nej, da $0 \notin [19.10, 19.11]$

Spørgsmål III.4 (8)

Antag, at vi istedet havde brugt den (uoplyste) teststørrelse `Q` til at teste, om der er en signifikant temperaturforskel mellem de to måneder. Hvad er de kritiske værdier ved brug af signifikansniveauet $\alpha = 0.01$?

1 ☐ ± 1.832

2 ☐ ± 1.960

3 ☐ ± 2.005

4 ☐ ± 2.398

5 ☐ ± 2.671

Spørgsmål III.5 (9)

Husejeren ønsker nu at teste om, der er en forskel mellem to givne dage, hvor der tages højde for timen på dagen. Han overvejer derfor en parret t-test.

Hvis X_i og Y_i betegner udfaldene fra de to stikprøver brugt i den parrede t-test, hvilket af følgende udsagn om antagelserne i den statistiske model er da korrekt?

Vi benytter notationen $\mathrm{V}[X_i] = \sigma_X^2$, $\mathrm{V}[Y_i] = \sigma_Y^2$ og $\mathrm{V}[X_i - Y_i] = \sigma_{X-Y}^2$ for varianserne, μ_X og μ_Y for middelværdierne i de to stikprøver og μ for forskellen i middelværdi.

1 ☐ $X_i \sim N(\mu, \sigma_X^2)$ og $Y_i \sim N(\mu, \sigma_Y^2)$, hvor begge er i.i.d. og uafhængige af hinanden

2 ☐ $X_i - Y_i \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$ og er i.i.d.

3 ☐ $X_i - Y_i \sim N(\mu, \sigma_{X-Y}^2)$ og er i.i.d.

4 ☐ $X_i - Y_i \sim N(0, \sigma_{X-Y}^2)$ og er i.i.d.

5 ☐ $X_i \sim N(\mu_X, \sigma_X^2)$ og $Y_i \sim N(\mu_Y, \sigma_Y^2)$, hvor begge er i.i.d. og uafhængige af hinanden

Fortsæt på side 7

Opgave IV

En energihandelsvirksomhed ønsker at blive klogere på elprisen i et givet område for en given periode. De henter data fra markedet og beregner den daglige elpris og relevante vejrvariable. Følgende variable er i datasættet:

- `Price`: Elektricitetsprisen på engrosmarkedet
- `Cloudcover`: Skydække (i %)
- `Humid`: Relativ luftfugtighed
- `Temperature`: Temperatur
- `Windspeed`: Vindhastighed

```
summary(lm(Price ~  Cloudcover + Humid + Temperature + Windspeed))

##
## Call:
## lm(formula = Price ~ Cloudcover + Humid + Temperature + Windspeed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30525 -0.04983  0.02637  0.07770  0.18326
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4419418  0.1080436   4.090 0.000139 ***
## Cloudcover   0.0003513  0.0006310   0.557 0.579901
## Humid        0.0003016  0.0010300   0.293 0.770754
## Temperature  0.0098091  0.0041229   2.379 0.020784 *
## Windspeed   -0.0529552  0.0127183  -4.164 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1112 on 56 degrees of freedom
## Multiple R-squared:  0.3757,Adjusted R-squared:  0.3311
## F-statistic: 8.427 on 4 and 56 DF,  p-value: 2.146e-05
```

Spørgsmål IV.1 (10)

Hvor mange dage er inkluderet i datasættet?

1 ☐ 51

2 ☐ 55

3 ☐ 56

4 ☐ 57

5 ☐ 61

Spørgsmål IV.2 (11)

Hvad er resultatet af det første backward selection step på modellen med signifikansniveau $\alpha = 0.05$ (både konklusionen og argumentet skal være korrekt)?

1 ☐ `Humid` skal fjernes, da $0.771 > 0.580 > 0.05$

2 ☐ `Windspeed` skal fjernes, da det har den højeste usikkerhed (uden at tage `Intercept` i betragtning)

3 ☐ `Windspeed` og `Temperature` skal fjernes, da $0.00011 < 0.05$ og $0.021 < 0.05$

4 ☐ `Humid` og `Cloudcover` skal fjernes, da $0.771 > 0.05$ og $0.580 > 0.05$

5 ☐ Ingen af variablene skal fjernes, da t-værdierne alle er numerisk større end $t_{\text{crit}} = 2.003$

Spørgsmål IV.3 (12)

Hvis man ser bort fra potentielle modelreduktioner, hvilken af de følgende konklusioner kan drages for markedet i den pågældende periode med det estimerede resultat?

1 ☐ Den estimerede middelværdi af prisen i perioden er 0.4419

2 ☐ Når temperaturen stiger så falder prisen, og når vindhastigheden stiger så stiger prisen

3 ☐ 99% prædiktionsintervallet for middelværdien af prisen har bredden $2 \cdot 0.111$

4 ☐ Modellen kan anvendes til at forudsige middelværdien af vindhastigheden i perioden

5 ☐ Modellen kan forklare 37.6% af den observerede variation i prisen i perioden

Fortsæt på side 9

Opgave V

Denne øvelse indeholder spørgsmål relateret til supermarkeder.

Spørgsmål V.1 (13)

Før i tiden indtastede kassemedarbejderne i supermarkedet priserne manuelt på kasseapparatet. Når medarbejderne var trætte, lavede de ofte fejl ved indtastning af priser. Antag, at for en bestemt situation lavede de tilfældigt en fejl på prisen for 5% af kunderne. Der antages uafhængighed af indtastningerne.

Hvad er sandsynligheden for, at 10 eller flere ud af 100 kunder får en fejl på prisen i denne situation?

1 ☐ 0.0015

2 ☐ 0.0043

3 ☐ 0.028

4 ☐ 0.063

5 ☐ 0.55

Spørgsmål V.2 (14)

I en undersøgelse af et supermarked antages ankomstraten af kunder at være 200 kunder/time i myldretiden. Kunder ankommer ifølge en Poisson-proces. Hvis der kommer mere end 250 kunder på en time, overskrides butikkens kapacitet. Hvad er sandsynligheden for, at butikkens kapacitet ikke overskrides i en tilfældigt udvalgt time i myldretiden?

1 ☐ 0.00028

2 ☐ 0.00061

3 ☐ 0.51879

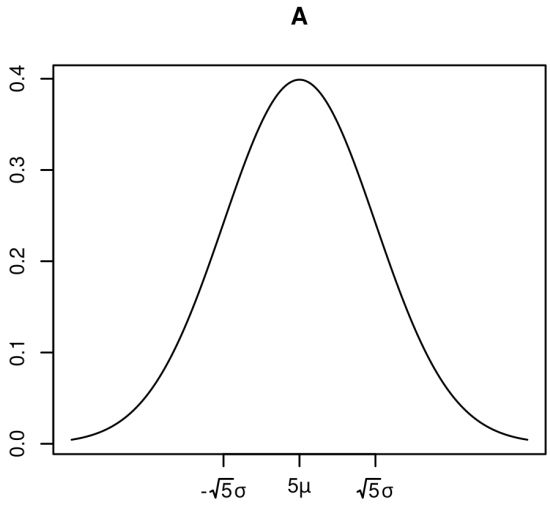
4 ☐ 0.92470

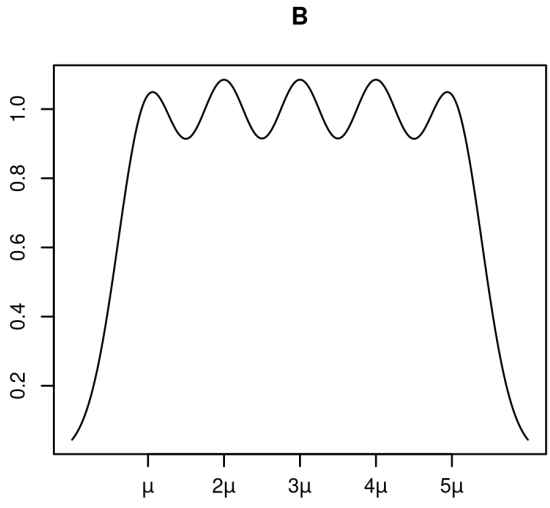
5 ☐ 0.99972

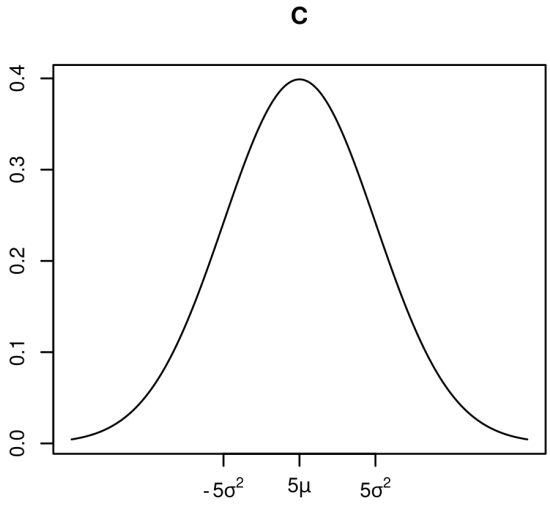
Spørgsmål V.3 (15)

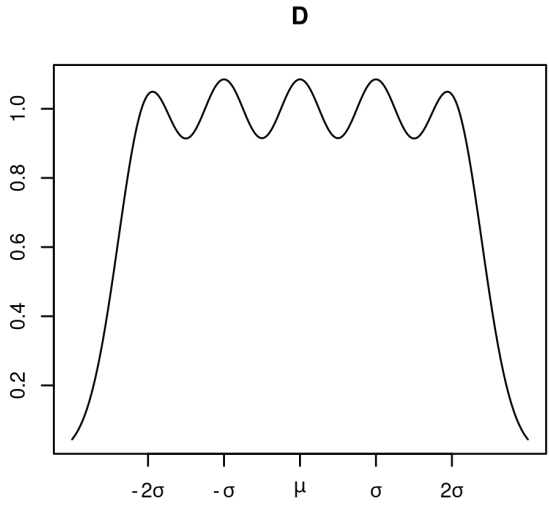
Lad $X \sim N(\mu, \sigma^2)$ angive den gennemsnitlige daglige omsætning i en bestemt butik. Butikken havde åbent 5 dage om ugen, og det kan antages, at de daglige omsætninger var uafhængige mellem dage.

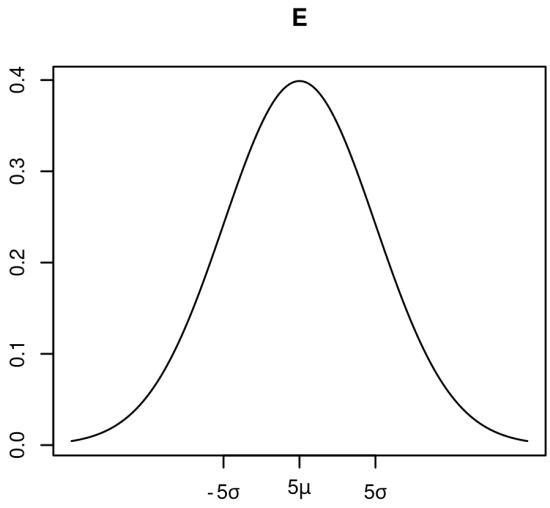
Et af følgende plots viser tætheden af den ugentlige omsætning. Hvilket?


A
0.0
0.1
0.2
0.3
0.4
-√5σ
5μ
√5σ


B
0.2
0.4
0.6
0.8
1.0
μ
2μ
3μ
4μ
5μ


C
0.0
0.1
0.2
0.3
0.4
-5σ²
5μ
5σ²


D
0.2
0.4
0.6
0.8
1.0
-2σ
-σ
μ
σ
2σ


E
0.0
0.1
0.2
0.3
0.4
-5σ
5μ
5σ

1 ☐ A

2 ☐ B

3 ☐ C

4 ☐ D

5 ☐ E

Fortsæt på side 13

Opgave VI

Lad X and Y være to uafhængige eksponentialfordelte stokastiske variable med rate hhv. 1.2 og 1.7.

Spørgsmål VI.1 (16)

Vi er interesseret i sandsynligheden for, at $X+Y$ er større end 3. Brug simulation til at vurdere, hvilken af nedenstående værdier er det korrekte restultat. Vi anbefaler, at du bruger mindst 10000 simulationer.

1 ☐ 0.078

2 ☐ 0.120

3 ☐ 0.344

4 ☐ 0.645

5 ☐ 0.920

Fortsæt på side 14

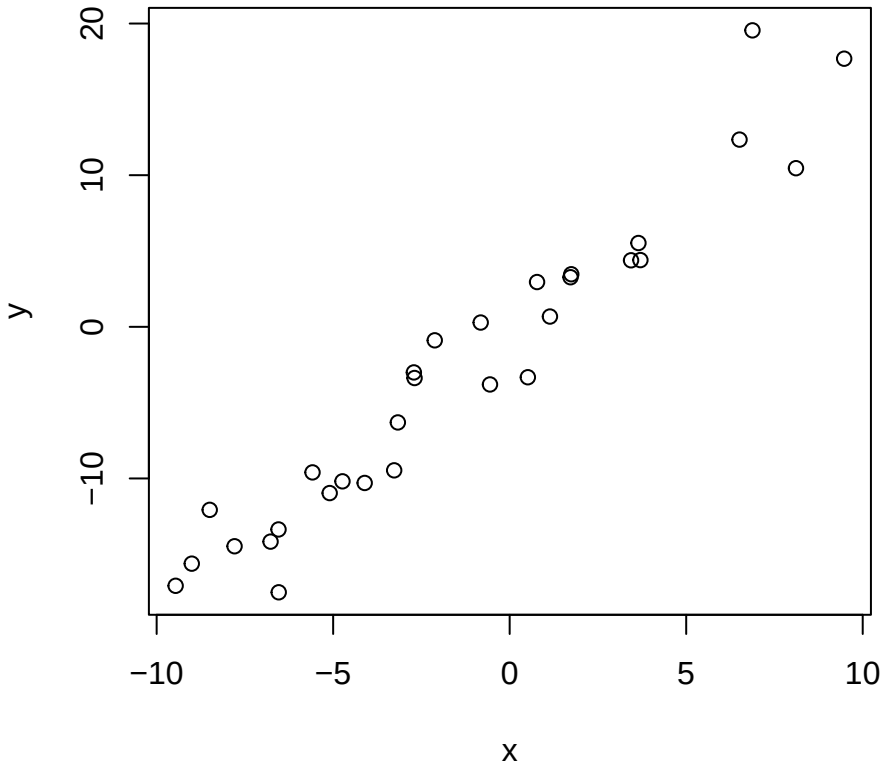
Opgave VII

Den simple lineære regressionsmodel er givet ved

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,$$

hvor $\varepsilon_i \sim N(0, \sigma^2)$ og uafhængige, $i = 1, \ldots, n$.

En stikprøve af de to parrede variable er gemt i R i vektorerne `x` og `y`. Et scatterplot af variablene er:



Den simple lineære regressionsmodel er blevet fittet, og resultatet er udskrevet herunder. Bemærk, at nogle af værdierne er blevet erstattet af bogstaver:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9599 -1.4571  0.1936  1.4127  7.2499
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.43369    0.49844   -0.87    0.392
## x               A       0.09284   19.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.636 on 28 degrees of freedom
## Multiple R-squared:  0.9342,Adjusted R-squared:  0.9319
## F-statistic: 397.8 on 1 and 28 DF,  p-value: < 2.2e-16
```

Spørgsmål VII.1 (17)

Hvilken af følgende værdier skal erstatte `A` i resultatet (tip: det kan være en hjælp at se på plottet)?

1 ☐ -0.73

2 ☐ 0.73

3 ☐ 1.85

4 ☐ 9.46

5 ☐ 20.15

Spørgsmål VII.2 (18)

Hvilket af følgende kald i R beregner bredden af 99% konfidensintervallet for β_0?

1 ☐ `2*qt(0.995,28) * 0.49844`

2 ☐ `2*qt(0.995,28) * 0.09284`

3 ☐ `qt(0.995,27) * 0.49844`

4 ☐ `qt(0.95,28) * 0.09284`

5 ☐ `qt(0.99,28) * 0.43369`

Fortsæt på side 16

Opgave VIII

De danske sundhedsmyndigheder (DSM) designer et forsøg for at undersøge alkoholforbruget blandt unge danskere. DSM ønsker specifikt at estimere andelen af unge danskere, der drikker mere end det maksimalt anbefalede antal genstande på en gennemsnitlig uge. DSM ønsker, at estimatet er inden for 0.01 af den faktiske andel med 95% sandsynlighed.

Spørgsmål VIII.1 (19)

Hvad er det minimale antal unge som skal inkluderes i forsøget for at opnå den ønskede præcision (vi vil her undlade at gøre nogen antagelse om den sande andel)?

1 ☐ 2401

2 ☐ 4147

3 ☐ 9604

4 ☐ 16588

5 ☐ 38415

Spørgsmål VIII.2 (20)

I et tidligere studie med 400 unge danskere har statistikere fra DSM accepteret nul-hypotesen $\mathcal{H}_0 : p = 0.25$ på et 10% signifikansniveau. Hvad er det mindste mulige estimat af andelen, som statistikerne kan have fundet i studiet?

1 ☐ $0.00\% = 0/400$

2 ☐ $20.75\% = 83/400$

3 ☐ $21.50\% = 86/400$

4 ☐ $23.25\% = 93/400$

5 ☐ $25.00\% = 100/400$

Fortsæt på side 17

Opgave IX

Som en del af et studie om adaptive læringsplatforme, deltog 47 frivillige studerende i afprøvningen af en ny undervisningsmetode for et helt semester. De studerendes resultater blev testet med en prætest før semesterstart og en posttest efter semesterets afslutning.

Prætest-scorer er gemt i `pretest`, og posttest-scorer er gemt i `posttest`. Begge er sorteret efter studienummer.

Spørgsmål IX.1 (21)

Følgende kode blev kørt:

```
sum(pretest)

## [1] 1620.042

quantile(pretest, probs = c(0.25, 0.5,  0.75))

##      25%      50%      75%
## 16.66667 30.00000 53.33333
```

Hvilket af følgende udsagn kan konkluderes om prætest-scorerne?

1 ☐ Gennemsnittet af prætest-scorerne er 30

2 ☐ Medianen af prætest-scorerne er 34.5

3 ☐ Interkvartilbredden (IQR) af prætest-scorerne er 36.7

4 ☐ Standardafvigelsen for prætest-scorerne er 16.7

5 ☐ Ingen af ovenstående

Spørgsmål IX.2 (22)

Vi ønsker at sammenligne de studerendes prætest- og posttestresultater, hvortil vi vil bruge den gennemsnitlige ændring i testscore (posttest minus prætest) som indikator.

Hvilket af følgende stykker R-kode udregner på korrekt vis et 95% konfidensinterval for dette ved brug af ikke-parametrisk bootstrapping?

1 ☐
```r
sim_mean_diff <- replicate(1000,
                           mean(sample(posttest, 20, replace = TRUE)) -
                           mean(sample(pretest, 20, replace = TRUE)))
quantile(sim_mean_diff, c(0.025, 0.975))
```

2 ☐
```r
sim_mean_diff <- replicate(1000,
                    mean(sample(posttest - pretest, 20, replace = TRUE)))
quantile(sim_mean_diff, c(0.025, 0.975))
```

3 ☐
```r
t.test(posttest, pretest , paired = FALSE, conf.level = 0.95)$conf.int
```

4 ☐
```r
t.test(posttest, pretest, paired = TRUE, conf.level = 0.95)$conf.int
```

5 ☐
```r
t.test(posttest, pretest, paired = TRUE, conf.level = 0.975)$conf.int
```

Spørgsmål IX.3 (23)

Som resultat af forrige spørgsmål fik man konfidensintervallet $[7.9, 17.2]$.

Hvilket af følgende udsagn kan vi konkludere?

1 ☐ Det gennemsnitlige posttest-resultat er signifikant større end det gennemsnitlige prætest-resultat på et 5% signifikansniveau

2 ☐ Det gennemsnitlige prætest-resultat er signifikant større end det gennemsnitlige posttest-resultat på et 5% signifikansniveau

3 ☐ Der er ikke en signifikant forskel på det gennemsnitlige prætest- og posttest-resultat på et 5% signifikansniveau

4 ☐ Der er en lineær sammenhæng mellem prætest- og posttest-resultater

5 ☐ Ingen af ovenstående

Fortsæt på side 19

Opgave X

Et hospital tog blodprøver fra 469 tilfældigt udvalgte mennesker fra forskellige aldersgrupper og screenede prøverne for et bestemt kemikalie. Resultaterne fra undersøgelserne er givet i Tabel 1 nedenfor:

Tabel 1	Aldersgr. 1	Aldersgr. 2	Aldersgr. 3	Aldersgr. 4	Total
Kemikalie ikke detekteret	17	28	21	15	81
Kemikalie detekteret	73	138	105	72	388
Total	90	166	126	87	469

Informationerne fra Tabel 1 kan indlæses i R med kodestumpen:

```
table1 <- matrix(c(17,28,21,15,73,138,105,72),nrow=2,byrow=TRUE)
```

Spørgsmål X.1 (24)

Under nulhypotesen, at sandsynligheden for at detektere kemikaliet i en blodprøve er den samme på tværs af alle aldersgrupperne, hvad er det forventede antal blodprøver, hvor kemikaliet ikke er blevet detekteret, fra folk i aldersgruppe 3?

1 ☐ 20.25

2 ☐ 21.76

3 ☐ 26.30

4 ☐ 28.67

5 ☐ 104.24

Spørgsmål X.2 (25)

Hvilken af følgende konklusioner er korrekt, hvis man tester nulhypotesen, at sandsynligheden for at detektere kemikaliet i en blodprøve er den samme på tværs af alle aldersgrupperne, på et 5% signifikansniveau (både forklaringen og konklusionen skal være korrekt)?

1 ☐ p-værdien er 0.025 og nul-hypotesen afvises derfor

2 ☐ p-værdien er 0.025 og nul-hypotesen accepteres derfor

3 ☐ p-værdien er 0.975 og nul-hypotesen afvises derfor

4 ☐ p-værdien er 0.975 og nul-hypotesen accepteres derfor

5 ☐ p-værdien er 0.975 og testen har derfor ikke en entydig konklusion

Spørgsmål X.3 (26)

Blodprøverne, hvor kemikaliet blev detekteret, blev yderligere inddelt som vist i Tabel 2 nedenfor:

Tabel 2	Aldersgruppe 1	Aldersgruppe 2	Aldersgruppe 3	Aldersgruppe 4	Total
Type A detekteret	35	64	42	20	161
Type B detekteret	30	60	55	45	190
Type C detekteret	8	14	8	7	37
Total	73	138	105	72	388

Informationerne fra Tabel 2 kan indlæses i R med kodestykket:

```
table2 <- matrix(c(35,64,42,20,30,60,55,45,8,14,8,7),nrow=3,byrow=TRUE)
```

Betragt nu kun blodprøverne, hvor kemikaliet blev detekteret. Hospitalspersonalet vil gerne teste, hvorvidt der er uafhængighed mellem typen af det detekterede kemikalie og aldersgruppen for den person, som afgav prøven, tilhører. Hvilken af følgende udtalelser er korrekt, hvis hospitalet benytter et konfidensniveau på 90%?

1 ☐ Den observerede teststørrelse er 10.177 og skal sammenlignes med χ_{crit}, hvor χ_{crit} er 90%-fraktilen i χ^2-fordelingen med 6 frihedsgrader

2 ☐ Den observerede teststørrelse er 10.177 og skal sammenlignes med χ_{crit}, hvor χ_{crit} er 90%-fraktilen i χ^2-fordelingen med 8 frihedsgrader

3 ☐ Den sædvanlige test afviser nulhypotesen om uafhængighed på det givne signifikansniveau

4 ☐ Den sædvanlige test er ugyldig, da nogle af de forventede værdier er mindre end 5

5 ☐ Sandsynligheden for at observere en teststørrelse mindre end 10.177 er 11.74% under nulhypotesen om uafhængighed

Fortsæt på side 21

Opgave XI

Spørgsmål XI.1 (27)

Bertil og Karin har indsamlet data til deres bachelorprojekt, og som en del af dette undersøger de sammenhængen mellem to variable, `height` og `time`.

De ønsker at anvende lineær regression, men er uenige om hvordan man korrekt tjekker modelantagelserne. Kun ét af nedenstående udsagn er korrekt. Hvilket?

1 ☐ Ikke-parametrisk bootstrapping af residualerne vil kunne afsløre, om antagelserne bag lineær regression er opfyldt

2 ☐ Et histogram af `height`-værdierne vil kunne afsløre, om normalitetsantagelsen er opfyldt

3 ☐ Værdien af variationskoefficienten (R^2) vil kunne afsløre, om linearitetsantagelsen er opfyldt

4 ☐ Et boxplot af `time`-værdierne vil kunne afsløre, om normalitetsantagelsen er opfyldt

5 ☐ Et QQ-plot af residualerne vil kunne afsløre, om normalitetsantagelsen er opfyldt

Fortsæt på side 22

Opgave XII

De følgende tider blev registreret af kvartmil-løbere (1/4 mile) og mil-løbere (1 mile) på et universitetsatletikhold (tiderne er i minutter). Observationerne læses ind i R ved:

```
quarter_mile_times <- c(0.92, 0.98, 1.04, 0.90, 0.99)
mile_times <- c(4.52, 4.35, 4.60, 4.70, 4.50)
```

Efter at have set disse løbetider kommenterede en af trænerne, at kvartmileløberne leverede mere ensartede tider.

Spørgsmål XII.1 (28)

Beregn standardafvigelserne og variationskoefficienterne (CV) for at opsummere variationen i dataene.

1 ☐ Kvartmil-løbere: $s = 0.0564, CV = 0.0584$.
Mil-løbere: $s = 0.1295, CV = 0.0286$.

2 ☐ Kvartmil-løbere: $s = 0.1295, CV = 0.0286$.
Mil-løbere: $s = 0.0564, CV = 0.0584$.

3 ☐ Kvartmil-løbere: $s = 0.0413, CV = 0.0584$.
Mil-løbere: $s = 0.1295, CV = 0.0286$.

4 ☐ Kvartmil-løbere: $s = 0.0564, CV = 0.0564$.
Mil-løbere: $s = 0.1295, CV = 0.0564$.

5 ☐ Kvartmil-løbere: $s = 0.0564, CV = 0.0413$.
Mil-løbere: $s = 0.0584, CV = 0.0564$.

Opgave XIII

En prøve på 12 af de bedst bedømte hoteller i USA har følgende antal værelser (`rooms`) og pris pr. nat for et dobbeltværelse (`cost`) (som læst i R).

```
rooms <- c(220, 727, 285, 273, 145, 213, 398, 343, 250, 414, 400, 700)
cost <- c(499, 340, 585, 495, 495, 279, 279, 455, 595, 367, 675, 420)
```

Spørgsmål XIII.1 (29)

Hvad er korrelationskoefficienten r for korrelationen mellem de to variable? Hvad fortæller det dig om forholdet mellem antallet af værelser og prisen per nat for et dobbeltværelse?

1 ☐ $r = -0.293$, en lille negativ korrelation. Højere pris pr. nat har en tendens til at være forbundet med større hoteller.

2 ☐ $r = -0.493$, en moderat negativ korrelation. Lavere pris pr. nat har en tendens til at være forbundet med større hoteller.

3 ☐ $r = -0.493$, en moderat negativ korrelation. Højere pris pr. nat har en tendens til at være forbundet med større hoteller.

4 ☐ $r = 0.791$, en stærk positiv korrelation. Højere pris pr. nat plejer at være forbundet med de større hoteller.

5 ☐ $r = -0.293$, en lille negativ korrelation. Lavere pris pr. nat har en tendens til at være forbundet med større hoteller.

Opgave XIV

Der blev indsamlet en stikprøve, og de tilhørende nøgletal blev beregnet.
Stikprøven er:

3, 6, 7, 0, 6, 13, 3, 7, 9, 15

Nøgletallene er (afrundet til to decimaler):

Nøgletal	Værdi
$\bar{x}$	6.9
s	4.56
s^2	20.77
Minimum	0
Q_1	3.75
Median	6.5
Q_3	8.5
Maximum	15
n	10

Spørgsmål XIV.1 (30)

Vi har dog mistanke om, at der er en fejl i et af nøgletallene, hvilket?

1 ☐ $\bar{x}$

2 ☐ s^2

3 ☐ Median

4 ☐ Q_1

5 ☐ Der er ingen fejl i nøgletallene.

Fortsæt på side 24

SÆTTET ER SLUT. God juleferie!