Danmarks Tekniske Universitet *Side 1 af 20 sider.*

Skriftlig prøve: 16. december 2018

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 14 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" svararket (6 separate sider) på CampusNet 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	II.1	III.1	III.2	III.3	IV.1	IV.2	V.1	V.2	V.3
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

Opgave	X.3	X.4	XI.1	XI.2	XII.1	XII.2	XIII.1	XIV.1	XIV.2	XIV.3
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Eksamenssættet består af 20 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.*

Opgave I

Ved analysen af en enkelt stikprøve antages 10 målinger at være uafhængige og stamme fra en normalfordeling med middelværdi μ og varians σ^2. Stikprøvens gennemsnit er $\overline{x} = 0.57$, mens stikprøvens standardafvigelse er $s = 0.32$.

Spørgsmål I.1 (1)

Hvilket af følgende er et standard 99% konfidensinterval for den teoretiske spredning σ?

1 ☐ $[0.20; 0.73]$

2 ☐ $0.57 \pm 1.96 \cdot 0.32$

3 ☐ $[0.22; 0.58]$

4 ☐ $[0.05; 0.34]$

5 ☐ $[0.03; 0.53]$

Opgave II

Vi vil gerne bestemme medianen for X_1/X_2, når X_1 og X_2 er uafhængige stokastiske variable, som begge er normalfordelte med middelværdi 1 og varians 1. Fordelingen af kvotienten er ikke triviel, og vi benytter derfor simulation til at bestemme et estimat og et konfidensinterval for medianen i fordelingen af X_1/X_2.

Spørgsmål II.1 (2)

Først simuleres 10000 medianer, som hver er medianen for 10000 kvotienter. Disse gemmes i R i vektoren `medians`:

```
ratio <- replicate(10000, rnorm(10000, mean = 1)/rnorm(10000, mean = 1))
medians <- apply(ratio, 2, median)
```

Derefter beregnes gennemsnittet og en række fraktiler for disse 10000 medianer:

```
mean(medians)

## [1] 0.6193
```

```r
quantile(medians, c(0.005, 0.025, 0.05, 0.5, 0.95, 0.975, 0.995), type = 2)
```

```
##   0.5%   2.5%     5%    50%    95%  97.5%  99.5%
## 0.5873 0.5949 0.5989 0.6193 0.6402 0.6443 0.6515
```

Hvilken af følgende valgmuligheder angiver et estimat for medianen af X_1/X_2 og et 95% konfidensinterval for denne median?

1 ☐ Estimat: 1.
95% konfidensinterval: $[1 - 1.96 \cdot 0.6193, 1 + 1.96 \cdot 0.6193]$.

2 ☐ Estimat: 0.6193.
95% konfidensinterval: $[0.5949, 0.6443]$.

3 ☐ Estimat: 1.
95% konfidensinterval: $[1 - 0.5949, 1 + 0.6443]$.

4 ☐ Estimat: 0.6193.
95% konfidensinterval: $[0.5873, 0.6515]$.

5 ☐ Estimat: 0.6193.
95% konfidensinterval: $[0.6193 - 0.5949, 0.6193 + 0.5949]$.

Opgave III

En normalfordelt population har middelværdi $\mu = 100$ og standardafvigelse $\sigma = 15$.

Spørgsmål III.1 (3)

Ved en tilfældig udtrækning, hvad er da sandsynligheden for at observationen er under 90?

1 ☐ 0.252

2 ☐ 0.482

3 ☐ 0.518

4 ☐ 0.631

5 ☐ 0.748

Spørgsmål III.2 (4)

Hvis der udtages en tilfældig stikprøve på $n = 10$ uafhængige observationer fra populationen, hvad er da sandsynligheden for, at stikprøvens gennemsnit er under 90?

1 ☐ 0.000783

2 ☐ 0.0175

3 ☐ 0.146

4 ☐ 0.252

5 ☐ 0.482

Spørgsmål III.3 (5)

Antag at der gentagne gange udtages en tilfældig stikprøve på n uafhængige observationer fra populationen, og at stikprøvevariansen S^2 beregnes i hver gentagelse. Hvad gælder der om S^2?

1 ☐ n^2S^2 er F-fordelt med $n-1$ og $n-2$ frihedsgrader.

2 ☐ S^2 er χ^2-fordelt med $n-1$ frihedsgrader.

3 ☐ $(n-1)S^2/\sigma^2$ er χ^2-fordelt med $n-1$ frihedsgrader.

4 ☐ S^2 er normalfordelt med middelværdi μ og varians σ^2/n^2.

5 ☐ S^2 har samme fordeling som $(Z-\sigma^2)/n$, hvor Z er standardnormalfordelt.

Opgave IV

10 personer har fået målt deres daglige energiindtag (i kJ). Målingerne i stikprøven fremgår af tabellen nedenfor:

Energiindtag (kJ):	8230	5470	7515	5260	6390	6180	6515	6805	7515	5640

Spørgsmål IV.1 (6)

Hvad er stikprøvens median?

1 ☐ 6390

2 ☐ 6515

3 ☐ (8230+5260)/2

4 ☐ (6390+6180)/2

5 ☐ (6390+6515)/2

Spørgsmål IV.2 (7)

Stikprøvegennemsnittet er $\overline{x} = 6552$, mens stikprøvespredningen er $s = 975.94$. Det antages, at det daglige energiindtag kan modelleres med en normalfordeling, og at observationerne er uafhængige og identisk fordelte. Hvad er p-værdien for det t-test, som tester hypotesen om, at det gennemsnitlige daglige energiindtag er 7725 kJ?

1 ☐ 0.4

2 ☐ 0.06

3 ☐ 0.04

4 ☐ 0.006

5 ☐ 0.004

Opgave V

Et ægtepar besøger samme restaurant flere gange om måneden og bestiller typisk et glas rødvin til maden. En dag beslutter de sig for at klage til ejeren. De mener, at én af tjenerne hælder mindre vin i glasset, end de betaler for. Ejeren iværksætter derfor et forsøg med tre af restaurantens tjenere for at undersøge, hvor meget de hælder op i vinglas, når de hælder på øjemål. Hver af de tre tjenere (her anonymiseret ved A, B og C) blev bedt om at hælde rødvin op i 20 vinglas, hvorefter indholdet i hvert glas blev målt. Data blev indlæst i R i to variable: `waiter`, som angiver hvilken tjener der hældte vinen op og `wine`, som angiver mængden af vin i glasset (i mL).

Følgende kode blev kørt i R for at analysere data:

```
anova(lm(wine ~ waiter))

## Analysis of Variance Table
##
## Response: wine
##           Df Sum Sq Mean Sq F value   Pr(>F)
## waiter     2 1043.4  521.71  6.9594 0.001976 **
## Residuals 57 4273.0   74.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål V.1 (8)

Hvad kan der konkluderes ud fra ovenstående R-output, når der benyttes et 5% signifikansniveau (både argument og konklusion skal være korrekt)?

1 ☐ Da den observerede F-teststørrelse er større end 0.95-fraktilen i $F(57, 2)$-fordelingen, er der signifikant forskel på middelindholdet af vin i glas hældt op af de tre forskellige tjenere.

2 ☐ Da p-værdien er større end 5%, er der ikke signifikant forskel på middelindholdet af vin i glas hældt op af de tre forskellige tjenere.

3 ☐ Da kvadratafvigelsessummen for residualerne, *SSE*, er over fire gange så stor som kvadratafvigelsessummen for grupperingen, *SS(Tr)*, er der for meget støj i data, til at det er meningsfuldt at lave envejs variansanalyse.

4 ☐ Da den observerede F-teststørrelse er større end 0.95-fraktilen i $F(2, 57)$-fordelingen er der signifikant forskel på middelindholdet af vin i glas hældt op af de tre forskellige tjenere.

5 ☐ Da p-værdien er mindre end 5%, er der ikke signifikant forskel på middelindholdet af vin i glas hældt op af de tre forskellige tjenere.

Spørgsmål V.2 (9)

Ejeren vil blandt andet lave en sammenligning mellem tjener A (den unge tjener, som ægteparret klagede over) og tjener B (en ældre tjener med mange års erfaring i branchen). Det oplyses, at tjener A i gennemsnit hældte 127 mL vin i hvert glas, mens tjener B i gennemsnit hældte 135 mL vin op. Bestem t-teststørrelsen for det post hoc parvise hypotesetest, der sammenligner middelindholdet af vin i glas hældt op af tjener A og tjener B.

1 ☐ $t_{obs} = -0.92$

2 ☐ $t_{obs} = -4.13$

3 ☐ $t_{obs} = -2.92$

4 ☐ $t_{obs} = -1.07$

5 ☐ $t_{obs} = -0.11$

Spørgsmål V.3 (10)

I tillæg til oplysningerne i det forrige spørgsmål oplyses det, at tjener C i gennemsnit hældte 136 mL op i hvert glas. Bestem den Bonferroni-korrigerede LSD ("least significant difference")-værdi, der benyttes til at lave alle mulige parvise sammenligninger mellem de tre tjenere og afgør, hvor der er signifikante forskelle (både LSD-værdi og konklusion skal være rigtige). Benyt signifikansniveauet $\alpha = 5\%$.

1 ☐ $\text{LSD}_{0.05/3} = 7$ mL, så der er signifikant forskel mellem middelindholdet af vin i glas hældt op af tjener B og C, men ingen signifikant forskel mellem tjener A og B eller mellem tjener A og C.

2 ☐ $\text{LSD}_{0.05/3} = 7$ mL, så der er signifikant forskel mellem middelindholdet af vin i glas hældt op af tjener A og B samt mellem tjener A og C, men ingen signifikant forskel mellem tjener B og C.

3 ☐ $\text{LSD}_{0.05/3} = 4$ mL, så der er signifikant forskel mellem middelindholdet af vin i glas hældt op af tjener A og B samt mellem tjener A og C, men ingen signifikant forskel mellem tjener B og C.

4 ☐ $\text{LSD}_{0.05/3} = 17$ mL, så der er signifikant forskel mellem middelindholdet af vin i glas hældt op af tjener A og B samt mellem tjener A og C, men ingen signifikant forskel mellem tjener B og C.

5 ☐ $\text{LSD}_{0.05/3} = 4$ mL, så der er signifikant forskel mellem middelindholdet af vin i glas hældt op af tjener B og C, men ingen signifikant forskel mellem tjener A og B eller mellem tjener A og C.

Opgave VI

En fjeder karakteriseres ved dens fjederkonstant, k. Når man strækker en fjeder ud gælder der jf. Hookes lov, at

$$F = -k \cdot x\,,$$

hvor x er den længde (i meter), som fjederen bliver forlænget med, og F er den anvendte kraft (i Newton). Man har lavet følgende seks observationer for en fjeder:

	1	2	3	4	5	6
x	0.22	0.24	0.26	0.28	0.30	0.32
F	-0.51	-0.85	-0.89	-1.59	-1.97	-2.06

Observationerne blev indlæst i to vektorer i R, hhv. `x` (længde) og `F` (kraft), hvorefter følgende model blev estimeret:

```
model1 <- lm(F ~ x)
```

Nedenfor ses output fra `summary(model1)`, hvor enkelte tal er erstattet af bogstaver:

```
##
## Call:
## lm(formula = F ~ x)
##
## Residuals:
##        1        2        3        4        5        6
## -0.04484 -0.04146  0.25365 -0.10667 -0.15758  0.09690
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2433     0.5483     A       C    **
## x           -16.8663     2.0148     B       D    **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1686 on 4 degrees of freedom
## Multiple R-squared:  0.946,Adjusted R-squared:  0.9325
## F-statistic: 70.08 on 1 and 4 DF,  p-value: 0.001114
```

Spørgsmål VI.1 (11)

Hvordan kan den statistiske model svarende til `model1` beskrives?

1 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, hvor Y_i beskriver den længde fjederen bliver forlænget med, når kraften x_i benyttes, og $\varepsilon_1, \ldots, \varepsilon_6$ antages at være uafhængige og identisk $N(0, \sigma^2)$-fordelte.

2 ☐ $Y_i = \beta_1 x_i + \varepsilon_i$, hvor Y_i beskriver den kraft der bruges til at forlænge fjederen med længden x_i, og $\varepsilon_1, \ldots, \varepsilon_6$ antages at være uafhængige og identisk $N(0, 1)$-fordelte.

3 ☐ $Y_i = \beta_1 x_i + \varepsilon_i$, hvor Y_i beskriver den længde fjederen bliver forlænget med, når kraften x_i benyttes, og $\varepsilon_1, \ldots, \varepsilon_6$ antages at være uafhængige og identisk $N(0, 1)$-fordelte.

4 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, hvor Y_i beskriver den længde fjederen bliver forlænget med, når kraften x_i benyttes, og $\varepsilon_1, \ldots, \varepsilon_6$ antages at være uafhængige og identisk $N(0, 1)$-fordelte.

5 ☐ $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, hvor Y_i beskriver den kraft der bruges til at forlænge fjederen med længden x_i, og $\varepsilon_1, \ldots, \varepsilon_6$ antages at være uafhængige og identisk $N(0, \sigma^2)$-fordelte.

Spørgsmål VI.2 (12)

Angiv et estimat for fjederkonstanten, k, med udgangspunkt i estimatet for hældningen i `model1`:

1 ☐ 0.5483

2 ☐ 3.2433

3 ☐ 2.0148

4 ☐ 16.8663

5 ☐ 5.2004

Spørgsmål VI.3 (13)

Det ønskes undersøgt, om modellens skæring (intercept) er signifikant forskellig fra nul. Angiv den relevante teststørrelse:

1 ☐ -8.371

2 ☐ 5.915

3 ☐ 0.004

4 ☐ 0.548

5 ☐ 0.169

Spørgsmål VI.4 (14)

Hvilken fordeling har den teststørrelse, som benyttes til at teste, om modellens hældning kan antages at være nul?

1 ☐ En t-fordeling med 6 frihedsgrader.

2 ☐ En standardnormalfordeling.

3 ☐ En F-fordeling med 6 frihedsgrader.

4 ☐ En normalfordeling med middelværdi nul og standardafvigelse 0.1686.

5 ☐ En t-fordeling med 4 frihedsgrader.

Spørgsmål VI.5 (15)

I en simpel lineær regression som ovenstående er estimatorerne for skærings- og hældningsparametrene ofte korrelerede. Hvornår er korrelationen nul?

1 ☐ Når standardafvigelsen af den afhængige variabel er 1.

2 ☐ Når hældningen er estimeret til 0.

3 ☐ Når gennemsnittet af den forklarende variabel er 0.

4 ☐ Når standardafvigelsen af den forklarende variabel er 1.

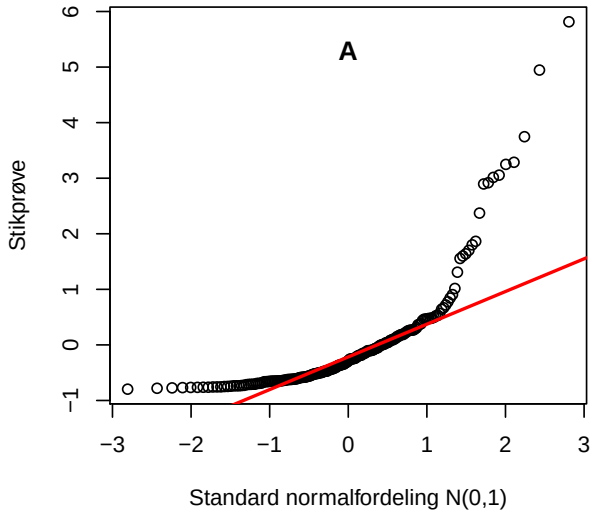
5 ☐ Når gennemsnittet af den afhængige variabel er 0.

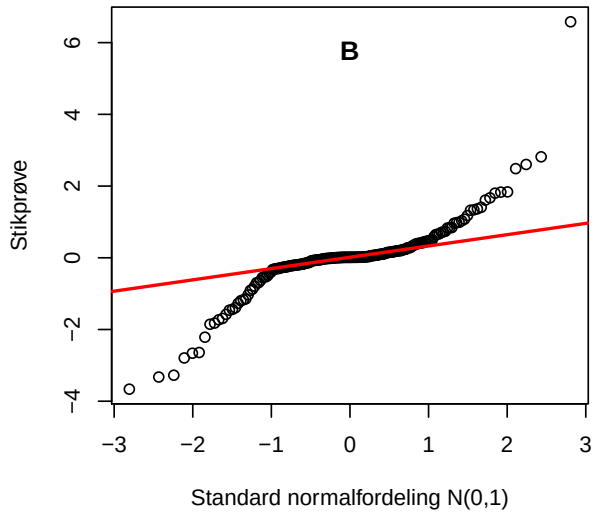
Opgave VII

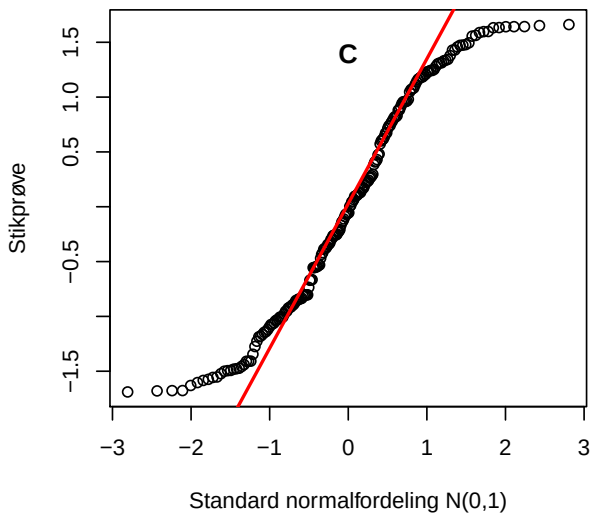
For at undersøge om data fra en enkelt stikprøve er log-normalfordelt kunne man sammenligne med en normalfordeling ved hjælp af et qq-plot. Hvis data er log-normalfordelt vil der (typisk) være færre små værdier og flere store værdier i data, sammenlignet med en normalfordeling der har samme middelværdi og varians som stikprøven.

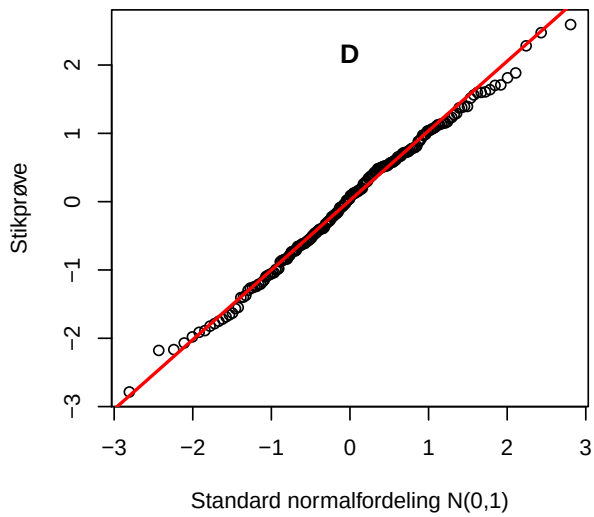
Spørgsmål VII.1 (16)

Nedenfor ses qq-plots, hvor fire forskellige stikprøver med middelværdi 0 og varians 1 hver bliver sammenlignet med en standardnormalfordeling. Lad $z_{0.25}$ og $z_{0.75}$ betegne hhv. første og tredje kvartil i standardnormalfordelingen, mens $q_{0.25}$ og $q_{0.75}$ er første og tredje kvartil i stikprøven. Den røde linje er trukket gennem punkterne $(z_{0.25}, q_{0.25})$ og $(z_{0.75}, q_{0.75})$. Hvilken stikprøve opfylder ovenstående beskrivelse af log-normalfordelt data?









1 ☐ A

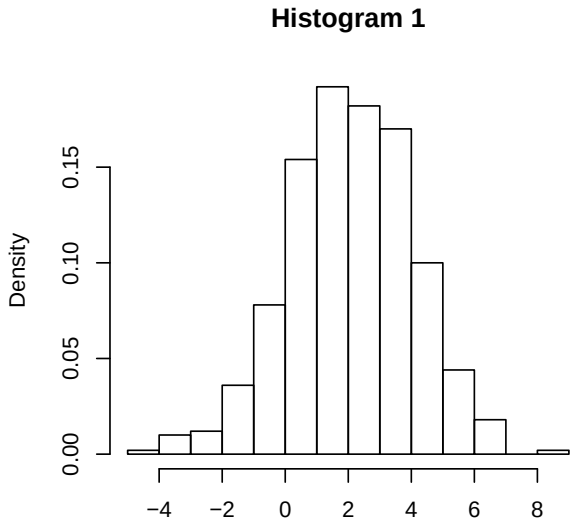
2 ☐ B

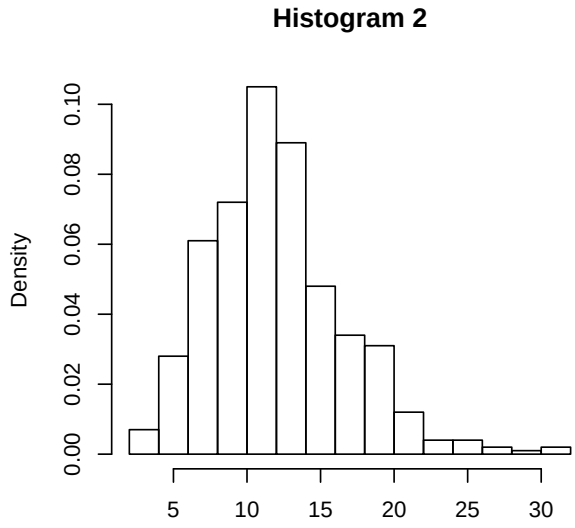
3 ☐ C

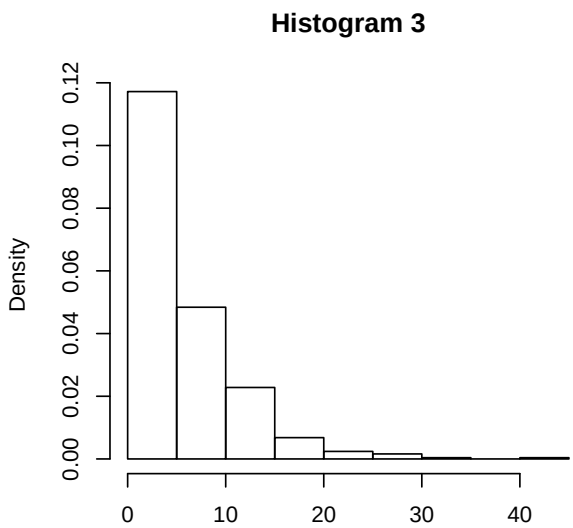
4 ☐ D

5 ☐ Ingen af ovenstående.

Opgave VIII







Spørgsmål VIII.1 (17)

Hvilke fordelinger er simuleret ovenfor? ($N(\mu, \sigma^2)$ refererer til normalfordelingen med middelværdi μ og varians σ^2, χ^2_a til χ^2-fordelingen med a frihedsgrader, $Exp(\beta)$ til eksponentialfordelingen med rate β).

1 ☐ 1: $N(0,4)$, 2: χ^2_{10}, 3: $Exp(1/5)$

2 ☐ 1: χ^2_4, 2: $N(2,4)$, 3: χ^2_1.

3 ☐ 1: $N(2,4)$, 2: χ^2_{12}, 3: $Exp(1/5)$

4 ☐ 1: $N(2,4)$, 2: $Exp(5)$, 3: χ^2_1

5 ☐ 1: $N(2,4)$, 2: χ^2_1, 3: $Exp(1/5)$

Opgave IX

To grupper rotter sættes på en diæt i løbet af opvæksten, og deres vægtforøgelse mellem dag 28 og dag 84 registreres. 10 rotter sættes på en diæt med højt proteinindhold, mens 7 rotter sættes på en diæt med lavt proteinindhold. De indsamlede data (vægtforøgelse i gram) fremgår af tabellen nedenfor, med den totale vægtforøgelse i hver gruppe angivet i nederste række:

	Højt proteinindhold	Lavt proteinindhold
	134	70
	146	118
	104	101
	119	85
	124	107
	161	132
	107	94
	83	
	113	
	129	
I alt	1220	707

Ud fra tallene i tabellen kan stikprøvevarianserne i de to grupper beregnes til $s^2_H = 495$ og $s^2_L = 425$, hvor H og L indikerer grupperne med henholdsvis højt og lavt proteinindhold. Det oplyses endvidere, at det sædvanlige test, for om rotter på diæt med henholdsvis højt og lavt proteinindhold har samme forventede tilvækst, har 13.7 frihedsgrader.

Spørgsmål IX.1 (18)

Hvilket af følgende valgmuligheder er korrekt (begge udsagn skal være rigtige)?

1 ☐ Rotter i gruppen med lavt proteinindhold i diæten tager mere på i vægt end rotter fra gruppen med højt proteinindhold i diæten. Forskellen er dog ikke signifikant ved signifikansniveau $\alpha = 0.05$.

2 ☐ Rotter i gruppen med højt proteinindhold i diæten tager mere på i vægt end rotter fra gruppen med lavt proteinindhold i diæten. Forskellen er signifikant ved signifikansniveau $\alpha = 0.05$.

3 ☐ Rotter i gruppen med højt proteinindhold i diæten tager mere på i vægt end rotter fra

gruppen med lavt proteinindhold i diæten. Forskellen er signifikant ved signifikansniveau $\alpha = 0.01$.

4 ☐ Rotter i gruppen med lavt proteinindhold i diæten tager mere på i vægt end rotter fra gruppen med højt proteinindhold i diæten. Forskellen er signifikant ved signifikansniveau $\alpha = 0.05$.

5 ☐ Rotter i gruppen med højt proteinindhold i diæten tager mere på i vægt end rotter fra gruppen med lavt proteinindhold i diæten. Forskellen er dog ikke signifikant ved signifikansniveau $\alpha = 0.05$.

Opgave X

Danmarks Statistik stiller tal om Danmark til rådighed på `www.statistikbanken.dk`, blandt andet tal om færdselsuheld. Følgende antal er udtaget derfra:

År	2010				2017			
Type	Alle		Sprit		Alle		Sprit	
Zone	By	Land	By	Land	By	Land	By	Land
Eneuheld	240	491	107	178	174	340	55	96
Øvrige	1779	988	161	84	1456	819	106	48

Tal under "alle" inkluderer alle uheld (inkl. spiritusuheld) og tal under "sprit" dækker udelukkende spiritusuheld.

Spørgsmål X.1 (19)

Angiv et 99% konfidensinterval for den samlede andel af spiritusuheld i 2010, hvor du bruger den relevante normalfordelingsapproksimation.

1 ☐ $0.848 \pm 2.58\sqrt{\frac{0.848}{3498}}$

2 ☐ $0.152 \pm 2.58\sqrt{\frac{0.848}{3498}}$

3 ☐ $0.848 \pm 1.96\sqrt{\frac{0.152 \cdot 0.848}{3498}}$

4 ☐ $0.848 \pm 2.58\sqrt{\frac{0.152}{3498}}$

5 ☐ $0.152 \pm 2.58\sqrt{\frac{0.152 \cdot 0.848}{3498}}$

Spørgsmål X.2 (20)

Antag at andelen af spiritusuheld i kategorien "eneuheld" er repræsentativ for den samlede andel af spirituskørsel. (Data fra "øvrige" uheld skal altså *ikke* benyttes i dette spørgsmål). Hvad kan der da ved brug af tallene i tabellen ovenfor og formuleringerne i bogens tabel 3.1 konkluderes om forskellen i spirituskørsel mellem årene 2010 og 2017?

1 ☐ Der er meget stærke beviser for, at der er sket et fald i andelen af spiritusuheld.

2 ☐ Der er svage beviser for, at der er sket et fald i andelen af spiritusuheld.

3 ☐ Der er få eller ingen beviser for, at der er sket en ændring i andelen af spiritusuheld.

4 ☐ Der er svage beviser for, at der er sket en stigning i andelen af spiritusuheld.

5 ☐ Der er meget stærke beviser for, at der er sket en stigning i andelen af spiritusuheld.

Spørgsmål X.3 (21)

Fra samme kilde findes også tal om hastighedsgrænsen på de vejstrækninger, hvor uheldene har fundet sted. Der er udtaget følgende data, som beskriver antallet af uheld i landzoner ved forskellige hastighedsgrænser i årene 2010 og 2017.

	2010	2017
0 til 50 km/t	54	58
50 til 100 km/t	1280	966
100 til 130 km/t	144	135

Hvad bliver resultatet af det sædvanlige test for, om fordelingen af uheld i hastighedsgrænseintervallerne er den samme i begge år (både konklusion og argument skal være korrekte)? Benyt signifikansniveauet $\alpha = 1\%$.

1 ☐ Der kan ikke påvises en signifikant forskel i fordelingen af hastighedsgrænserne mellem de to år, da p-værdien er over signifikansniveauet.

2 ☐ Der kan påvises en signifikant forskel i fordelingen af hastighedsgrænserne mellem de to år, da p-værdien er over signifikansniveauet.

3 ☐ Der kan påvises en signifikant forskel i fordelingen af hastighedsgrænserne mellem de to år, da p-værdien er under signifikansniveauet.

4 ☐ Der kan ikke påvises en signifikant forskel i fordelingen af hastighedsgrænserne mellem de to år, da p-værdien er under signifikansniveauet.

5 ☐ Ingen af ovenstående udsagn er korrekte.

Spørgsmål X.4 (22)

I forbindelse med det sædvanlige test for, om fordelingen af hastighedsgrænserne er den samme i de to år, ønskes følgende besvaret: Hvad bliver den estimerede andel af uheld på veje med hastighedsgrænser fra 50 til 100 km/t i år 2017 under nulhypotesen?

1 ☐ $(58 + 966 + 135)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.440$

2 ☐ $(966)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.366$

3 ☐ $(966)/(58 + 966 + 135) = 0.833$

4 ☐ $(1280 + 966)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.852$

5 ☐ $(54 + 58 + 144 + 135)/(54 + 1280 + 144 + 58 + 966 + 135) = 0.148$

Opgave XI

Nedenfor ses en stikprøve med 20 uafhængige observationer, som er indlæst i R i vektoren `x`:

```
x <-
  c(13, 12, 9, 7, 12, 15, 12, 10, 6, 13, 7, 13, 19, 12, 6, 4, 15, 16, 11, 18)
```

Data stammer ikke umiddelbart fra en kendt fordeling, men vi interesserer os for populationens middelværdi og usikkerheden på estimatet af denne.

Spørgsmål XI.1 (23)

Hvad er stikprøvens gennemsnit $\bar{x}$ og varians s^2 (begge størrelser skal være korrekte)?

1 ☐ $\bar{x} = 11.2$ og $s^2 = 16.7$.

2 ☐ $\bar{x} = 11.5$ og $s^2 = 16.7$.

3 ☐ $\bar{x} = 11.2$ og $s^2 = 4.1$.

4 ☐ $\bar{x} = 11.5$ og $s^2 = 4.1$.

5 ☐ $\bar{x} = 11.5$ og $s^2 = 16.7^2$.

Spørgsmål XI.2 (24)

Vi udfører nu en resampling af `x` for at få en idé om usikkerheden på gennemsnittet. Der trækkes derfor 200 gentagelser (stikprøver af størrelse 20) med tilbagelægning fra de 20 observationer i `x`. Herefter tages gennemsnittet af hver af de 200 gentagelser. R-koden for den operation er:

```r
apply(replicate(200, sample(x, replace = TRUE)), 2, mean)
```

Nedenfor er vist de 10 største og 10 mindste gennemsnit fra de 200 gentagelser.

mindste	9.00	9.65	9.65	9.80	9.90	9.95	10.00	10.00	10.00	10.05
største	12.95	12.95	12.95	13.00	13.05	13.10	13.10	13.10	13.15	13.40

Ud fra ovenstående resultater og ved at benytte bogens definition af fraktiler ("`type = 2`" i R), hvilket af følgende er et 95% bootstrap konfidensinterval for populationens middelværdi?

1 ☐ $[10.05; 12.95]$

2 ☐ $[9.00; 13.40]$

3 ☐ $[9.80; 13.10]$

4 ☐ $[9.65; 13.10]$

5 ☐ $[9.925; 13.075]$

Opgave XII

Ved afholdelsen af en mindre festival skal der regnes på toiletforholdene. Der skal nemlig bestilles mobile toiletter, så kapaciteten er rigelig men ikke for stor, da sidstnævnte ville skabe mere rengøringsarbejde og koste flere penge.

Det antages, at der i gennemsnit er 150 gæster i timen, der skal benytte toilettet, og at deres ankomst følger en Poisson fordeling. Desuden antages det, at hvert toilet kan betjene 20 gæster i timen.

Spørgsmål XII.1 (25)

Antag at der bestilles 10 toiletter. Hvad er da sandsynligheden for, at der ankommer flere gæster til toiletterne i en tilfældig udvalgt time, end der er kapacitet til?

1 ☐ 0.0042%

2 ☐ 2.3%

3 ☐ 11%

4 ☐ 24%

5 ☐ 99%

Spørgsmål XII.2 (26)

En gruppe DTU-studerende har besluttet sig for at hjælpe mindre festivaller med at optimere deres logistiske forhold. De studerende har blandt andet indsamlet data om brugen af toiletter på mindre festivaller. En undersøgelse af disse data viser, at en bedre model kan laves til at repræsentere antallet af gæster, som skal på toilettet i en tilfældig udvalgt time. Dette antal kan modelleres ved en eksponentialfordeling med middelværdi $\frac{\text{"antal gæster"}}{10}$, hvor "antal gæster" er det totale antal gæster på festivallen. I dette spørgsmål skal denne nye model benyttes.

Der skal nu regnes på en festival med 1500 gæster. Hvor mange toiletter skal der mindst bestilles for at sikre at sandsynligheden for, at ikke alle kan komme på toilettet, er under 2 % i en tilfældig udvalgt time (angivet som kald i R)? Det antages fortsat, at hvert toilet kan betjene 20 gæster i timen.

1 ☐ `ppois(20, lambda = 15) * 20`

2 ☐ `qpois(0.98, lambda = 1500/10) / 20`

3 ☐ `qexp(0.98, rate = 10/15)`

4 ☐ `qexp(0.98, rate = 10/1500) / 20`

5 ☐ `qexp(0.98, rate = 10/1500) * 20`

Opgave XIII

Nedenfor ses en lille stikprøve med 5 uafhængige observationer:

Observationer:	11.8071067	-1.7913888	-9.1872410	-4.4860901	-0.2324924

Spørgsmål XIII.1 (27)

Hvilken af følgende svarmuligheder er den eneste, der kan være korrekt?

1 ☐ Observationerne kan umuligt stamme fra en normalfordeling med middelværdi 0 og varians 10^2.

2 ☐ Det er muligt, at observationerne stammer fra en uniform fordeling med parametre -9 og 12.

3 ☐ Det er muligt, at observationerne stammer fra en t-fordeling med 1 frihedsgrad.

4 ☐ Det er muligt, at observationerne stammer fra en F-fordeling med 1 og 2 frihedsgrader.

5 ☐ Det er muligt, at observationerne stammer fra en eksponentialfordeling med rate 1.

Opgave XIV

Efterhånden som vindenergi udgør en øget andel af energiproduktionen i Danmark og resten af Europa, bliver præcise forudsigelser mere afgørende. Forudsigelse af vindenergi afhænger naturligvis af vejrudsigten, men også estimationsmetode og modelstruktur har indflydelse. Tabellen herunder viser et udsnit af et datasæt, der angiver den gennemsnitlige ugentlige forudsagte vindenergiproduktion for en vindmøllepark (målt som andel af installeret effekt), for forskellige prædiktionsmodeller (m1, ..., m5).

Uge	m1	m2	m3	m4	m5
1	0.6039	0.6232	0.6083	0.5751	0.6232
2	0.5143	0.5301	0.5049	0.4644	0.4850
3	0.5551	0.5603	0.5415	0.5091	0.5219
4	0.5396	0.5393	0.5766	0.4697	0.5245
⋮	⋮	⋮	⋮	⋮	⋮

Nedenfor er data indlæst i R. Vektoren `prediction` indeholder de gennemsnitlige ugentlige forudsigelser, `model` angiver hvilken af de fem prædiktionsmodeller, der er benyttet, og `week` angiver ugenummer.

Man ønsker nu at undersøge, om de 5 forskellige modeller (m1, ..., m5), kan antages at give de samme forudsigelser i middel (på ugeniveau), eller om der er signifikant forskel.

For at undersøge hypotesen har man opstillet følgende model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

hvor α_i beskriver effekten af model i, og β_j beskriver effekten af uge j.

Spørgsmål XIV.1 (28)

Under de sædvanlige antagelser, hvilket af følgende udsagn er da korrekt?

1 ☐ $\alpha_i + \beta_j = 0$ for alle kombinationer af i og j.

2 ☐ ϵ_{ij} er uafhængige og normalfordelte med middelværdi 0 og en varians, der afhænger af α_i og β_j.

3 ☐ Y_{ij} er uafhængige og normalfordelte med samme varians for alle kombinationer af i og j.

4 ☐ $\sum_i \alpha_i = \sum_j \beta_j = \mu$.

5 ☐ Y_{ij} er uafhængige og identisk fordelte for alle kombinationer af i og j.

Spørgsmål XIV.2 (29)

For at undersøge hypotesen om, at der ikke er forskel (i middel) på forudsigelserne har man kørt nedenstående R-kode. Bemærk at en del af resultaterne er fjernet, og at enkelte tal er erstattet med bogstaver.

```
anova(lm(prediction ~ model + factor(week)))

Analysis of Variance Table

Response: pred
             Df  Sum Sq   Mean Sq F value    Pr(>F)
model         4 0.01056 0.0026391  7.3754  5.389e-05
factor(week)  A 0.33946 0.0199684 55.8051  < 2.2e-16
Residuals     B 0.02433 0.0003578
```

Hvor mange uger indgår der i datasættet? (På grund af afrunding i R-outputtet vil den relevante beregning resultere i et decimaltal, som skal afrundes korrekt til nærmeste hele tal).

1 ☐ 18

2 ☐ 68

3 ☐ 56

4 ☐ 17

5 ☐ 55

Spørgsmål XIV.3 (30)

Betragt igen R-outputtet fra forrige spørgsmål og brug signifikansniveauet $\alpha = 0.01$. Er der signifikant forskel på de fem modeller $m_1, \ldots, m_5$, når den statistiske model tager højde for forskellen mellem uger (både konklusion og argument skal være korrekte)?

1 ☐ Nej, da $0.0106 > 0.01$.

2 ☐ Ja, da $5.389 \cdot 10^{-5} < 0.01$.

3 ☐ Nej, da $0.020 > 0.01$.

4 ☐ Ja, da $0.0026 < 0.01$.

5 ☐ Ja, da $0.024 > 0.01$.

Eksamenssættet er slut. God juleferie!