Danmarks Tekniske Universitet *Side 1 af 28 sider.*

Skriftlig prøve: 19. Dec 2020

Kursus navn og nr.: **Introduction to Statistics (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 10 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	II.2	II.3	II.4	II.5	III.1	III.2	IV.1	IV.2
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

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

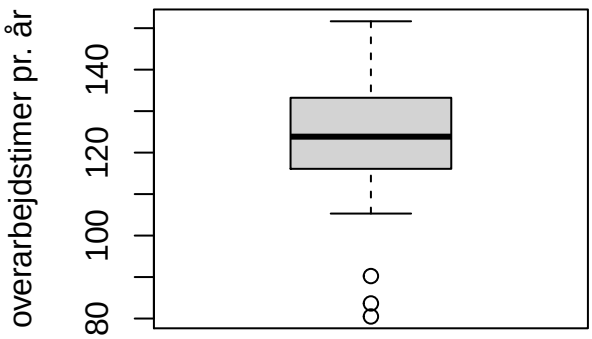
Eksamenssættet består af 28 sider.

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 byafdeling har indført et kvalitetsforbedringsprogram, der giver medarbejderne mulighed for at få kompensation for overarbejdstimer, når de deltager i møder. Det samlede antal overarbejdstimer pr. år for 36 ansatte visualiseres i nedenstående boxplot.



Spørgsmål I.1 (1)

Hvilken af følgende udsagn er korrekt?

1 ☐ $IQR = Q1 - Q3 \approx 17$ timer

2 ☐ $IQR = Q3 - Q1 \approx 17$ timer

3 ☐ $IQR = Q4 - Q1 \approx 48$ timer

4 ☐ IQR kan ikke bestemmes, fordi boxplot indeholder tre outliers

5 ☐ $IQR = Q3 - Q1 \approx 48$ timer

Fortsæt på side 3

Opgave II

Tabellen herunder viser antallet af personer, som er testet positiv for coronavirus, der blev indlagt på hospitaler i Danmark på 3 forskellige datoer i foråret 2020. Tabellen viser endvidere antallet af personer, der også var på en intensivafdeling (ICU).

Date	ICU	Admitted
April 30	62	255
April 10	113	433
March 20	37	153

Spørgsmål II.1 (2)

Baseret på tallene i tabellen, hvad er det sædvanlige 95% konfidensinterval for sandsynligheden for, at hvis du indlægges, er du også indlagt på en intensivafdeling? Antag, at modelantagelserne er opfyldt.

1 ☐ $[0.72, 0.78]$

2 ☐ $[0.22, 0.28]$

3 ☐ $[0.18, 0.22]$

4 ☐ $[0.16, 0.35]$

5 ☐ $[0.12, 0.28]$

Spørgsmål II.2 (3)

For at undersøge udviklingen over tid sammenlignes tallene fra 30. april og 20. marts. Med nulhypotesen om at andelen af patienter i ICU kan være ens på de to datoer, hvad er p-værdien og konklusionen givet signifikansniveau $\alpha = 0.05$? (Både konklusion og argument skal være korrekt)

1 ☐ p-værdi=0.476 og forskellen er signifikant.

2 ☐ p-værdi=0.029 og forskellen er ikke signifikant.

3 ☐ p-værdi=0.976 og forskellen er ikke signifikant.

4 ☐ p-værdi=0.024 og forskellen er signifikant.

5 ☐ p-værdi=0.060 og forskellen er ikke signifikant.

Spørgsmål II.3 (4)

Fordelingen af patienter over forskellige regioner undersøges nu. Tabellen herunder viser antallet af personer indlagt på hospitaller på forskellige datoer i de 5 regioner i Danmark. Vi antager her, at den samme person ikke er indlagt på mere end 1 dato.

Dato	Nordjylland	Midtjylland	Syddanmark	Hovedstaden	Sjælland	All DK
April 30	13	33	12	144	53	255
April 16	21	54	35	183	60	353
April 2	32	77	85	251	86	531
Marts 18	10	16	12	64	27	129
Total	76	180	144	642	226	1268

Vi vil nu undersøge, om andelen af indlagte patienter i de forskellige regioner er den samme over tid (nul-hypotesen), eller om den ændrer sig. Formelt kan dette skrives som

$$\mathrm{H}_0: \quad p_{ij} = p_i$$

for alle i.

Under nulhypotesen, hvad er bidraget til teststørrelsen for "Nordjylland " den 18. marts?

1 ☐ 7.73

2 ☐ 0.59

3 ☐ 5.14

4 ☐ 0.67

5 ☐ 10

Spørgsmål II.4 (5)

Test størrelsen er beregnet til $\chi^2_{obs} = 29$. Givet et signifikansniveau $\alpha = 0.05$, hvad er p-værdien og konklusionen for den tilsvarende hypotesetest? (Både konklusion og argument skal være korrekt)

1 ☐ p-værdi=0.0012 og der er en signifikant forskel

2 ☐ p-værdi=0.0099 og der er ikke en signifikant forskel

3 ☐ p-værdi=0.024 og der er en a signifikant forskel

4 ☐ p-værdi=0.088 og der er ikke en signifikant forskel

5 ☐ p-værdi=0.0039 og der er en signifikcant forskel

Spørgsmål II.5 (6)

Hvis vi på en given dag antager, at 4% af befolkningen er inficeret med en virus, hvor mange mennesker skal så testes tilfældigt for at få en fejlmargin på maksimalt 1% ved brug af signifikansniveau $\alpha = 0.05$?

1 ☐ 1039

2 ☐ 1476

3 ☐ 369

4 ☐ 9603

5 ☐ 6764

Fortsæt på side 6

Opgave III

Løn for akademisk arbejde over ni måneder i 2008-09 for professorer på et U.S.-universitet vurderes. Dataene inkluderer lønninger til 125 mandlige professorer, der arbejder i anvendte instituter (i US dollars). Vi vil finde ud af, om lønnen afhænger af antal arbejdsår siden at have fået en ph.d. og års tjeneste.

Spørgsmål III.1 (7)

En første multipel lineær regressionsmodel blev etableret. Modeloversigten er angivet nedenfor. Antag, at modelantagelserne er opfyldt!

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd + yrs.service, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72479 -20472   -288  16051  92778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   130213.8     6956.5  18.718   <2e-16 ***
## yrs.since.phd   -304.2      430.1  -0.707    0.481
## yrs.service      529.3      378.5   1.398    0.165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26450 on 122 degrees of freedom
## Multiple R-squared:  0.02085,Adjusted R-squared:  0.004803
## F-statistic: 1.299 on 2 and 122 DF,  p-value: 0.2765
```

Hvilken af de følgende udsagn er korrekt i betragtning af et signifikansniveau $\alpha = 0.05$? (Både konklusion og argument skal være korrekt)

1 ☐ Professor `salary` afhænger af `yrs.since.phd` og `yrs.service`, fordi begge p-værdier er større end 0.05.

2 ☐ Professor `salary` afhænger IKKE af `yrs.since.phd` og `yrs.service`, fordi begge p-værdier er større end 0.025.

3 ☐ Vi har ikke tilstrækkelig information til at konkludere om relation mellem Professor `salary` og `yrs.since.phd` og `yrs.service`.

4 ☐ Professor `salary` afhænger af `yrs.since.phd` og `yrs.service`, fordi begge p-værdier er mindre end 0.5.

5 ☐ Professor `salary` afhænger IKKE af `yrs.since.phd` og `yrs.service`, fordi begge p-værdier er større end 0.05.

Spørgsmål III.2 (8)

Backwards model selection blev udført for den multiple lineære regressionsmodel ovenfor, hvilket resulterede i følgende R-output:

```
##
## Call:
## lm(formula = salary ~ yrs.service, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73189 -20581     29  15226  92951
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 126901.7     5133.7  24.719   <2e-16 ***
## yrs.service    307.7      212.0   1.451    0.149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26400 on 123 degrees of freedom
## Multiple R-squared:  0.01684,Adjusted R-squared:  0.008846
## F-statistic: 2.107 on 1 and 123 DF,  p-value: 0.1492
```

```
##
## Call:
## lm(formula = salary ~ 1, data = sal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65959 -19018   -693  16858  98027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   133518       2372    56.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26510 on 124 degrees of freedom
```

Hvilken R-kode resulterer i det korrekte 95% konfidensinterval for gennemsnittet af Professor lønnen?

1 ☐ `133518 + c(-1, 1) * qt(0.95, 124) * 2372`

2 ☐ `133518 + c(-1, 1) * qt(0.975, 123) * 2372`

3 ☐ `133518 + c(-1, 1) * qt(0.975, 124) * 2372`

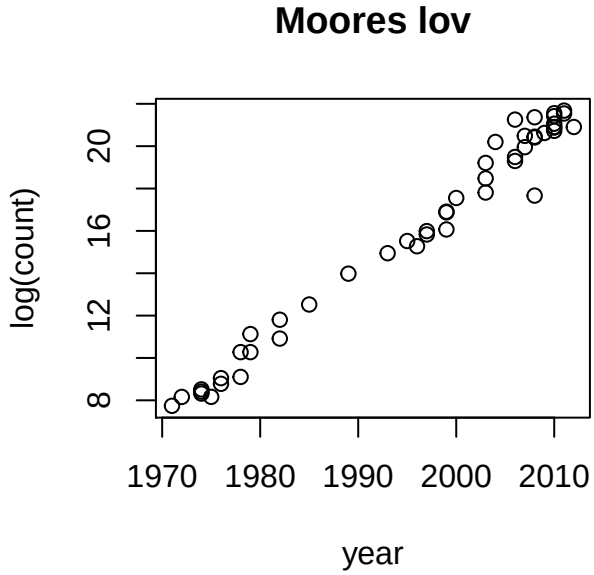
4 ☐ `126902 + c(-1, 1) * qt(0.975, 124) * 5134`

5 ☐ `130214 + c(-1, 1) * qt(0.95, 124) * 6957`

Fortsæt på side 10

Opgave IV

Moores lov handler om den observation, at antallet af transistorer i et tæt integreret kredsløb fordobles cirka hvert andet år. Observationen er opkaldt efter Gordon Moore, medstifter af Fairchild Semiconductor. I figuren nedenfor er transistorantallet transformeret ved hjælp af den naturlige logaritme og plottet mod år.




```
##
## Call:
## lm(formula = log(count) ~ year, data = moore)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60701 -0.26843 -0.01245  0.35038  1.67737
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.786e+02  1.414e+01  -48.01        ?
## year         3.481e-01  7.083e-03       ?   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 0.6762 on 46 degrees of freedom
## Multiple R-squared:  0.9813,Adjusted R-squared:  0.9809
## F-statistic:  2415 on 1 and 46 DF,  p-value: < 2.2e-16
```

Spørgsmål IV.1 (9)

Beregn den teststørrelse, som mangler i modeloversigten ovenfor (manglende værdier er erstattet af spørgsmålstegn i tabellen). Hvilket af følgende svar er rigtigt?

1 ☐ $t_{\text{obs}} = 0.02$

2 ☐ $t_{\text{obs}} = 12.25$

3 ☐ $t_{\text{obs}} = 0.49$

4 ☐ $t_{\text{obs}} = 49.15$

5 ☐ $t_{\text{obs}} = 12.49$

Spørgsmål IV.2 (10)

Vi vil teste hypotesen $H_0 : \beta_0 = 0$, hvor β_0 repræsenterer modellens intercept. Hvilket af de følgende udsagn er korrekt (antag $\alpha = 0.05$)? (Både argumentation og konklusion skal være korrekte!)

1 ☐ Vi sammenligner den absolutte værdi af den tilsvarende teststørrelse $|t_{obs}| = 48.01$ med den kritiske t-værdi, $t_{crit} = 1.96$. Vi afviser H_0, fordi $|t_{obs}| > t_{crit}$.

2 ☐ Vi sammenligner den absolutte værdi af den tilsvarende teststørrelse $|t_{obs}| = 48.01$ med den kritiske t-værdi, $t_{crit} = 2.01$. Vi afviser H_0, fordi $|t_{obs}| > t_{crit}$.

3 ☐ Vi sammenligner den absolutte værdi af den tilsvarende teststørrelse $|t_{obs}| = 48.01$ med den kritiske t-værdi, $t_{crit} = 1.68$. Vi afviser H_0, fordi $|t_{obs}| > t_{crit}$.

4 ☐ Vi sammenligner den absolutte værdi af den tilsvarende teststørrelse $|t_{obs}| = 48.01$ med den kritiske t-værdi, $t_{crit} = 2.01$. Vi accepterer H_0, fordi $|t_{obs}| > t_{crit}$.

5 ☐ Vi sammenligner den absolutte værdi af den tilsvarende teststørrelse $|t_{obs}| = 48.01$ med den kritiske t-værdi, $t_{crit} = 1.96$. Vi accepterer H_0, fordi $|t_{obs}| > t_{crit}$.

Spørgsmål IV.3 (11)

I henhold til den lineære model ovenfor, hvad er den forventede stigning i antallet af transistorer fra 2010 til 2015?

1 ☐ $\ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015) - \ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010)$

2 ☐ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015 + 6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010}$

3 ☐ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015} - e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010}$

4 ☐ $\ln(-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015 + 6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010)$

5 ☐ $e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2010} - e^{-6.786 \cdot 10^2 + 3.481 \cdot 10^{-1} \cdot 2015}$

Fortsæt på side 12

Opgave V

Spørgsmål V.1 (12)

Man er interesseret i at bestemme massefylden af en væske. For at gøre dette måles væskens masse, m, og volumenet, V. Massefylden er angivet ved

$$\rho = \frac{m}{V}$$

Hvad er præcisionen (standardafvigelse, σ_ρ) for den udregnede massefylde, hvis massen og volumenet kan måles med en præcision henholdsvis $\sigma_m = 0.2$ og $\sigma_V = 0.4$? Antag at masse- og volumenmålinger er uafhængige og normalfordelte.

1 ☐ $\sigma_\rho \approx \frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})$

2 ☐ $\sigma_\rho \approx \sqrt{\frac{1}{V^2}(0.2^2 + \frac{0.4^2 m^2}{V^2})}$

3 ☐ $\sigma_\rho \approx \frac{1}{V^2}(0.4^2 + \frac{0.2^2 m^2}{V^2})$

4 ☐ $\sigma_\rho \approx \frac{0.4^2}{V^2} + \frac{0.2^2 m^2}{V^4}$

5 ☐ $\sigma_\rho \approx \sqrt{\frac{0.4^2}{V^2} + \frac{0.2^2 m^2}{V^4}}$

Spørgsmål V.2 (13)

Lad X_i være en stokastisk variabel. Følgende kode køres i R for at trække 100 tilfældige tal X_i fra en given fordeling.

```
x <- rnorm(100)^2 + rnorm(100)^2 + rnorm(100)^2
```

Hvilket af følgende udsagn er korrekt?

1 ☐ X_i følger en χ^2-fordeling med 1 frihedsgrad.

2 ☐ X_i følger en standard normal fordeling med middelværdi 0 og varians 1.

3 ☐ X_i følger en χ^2-fordeling med 2 frihedsgrader.

4 ☐ X_i følger en χ^2-fordeling med 3 frihedsgrader.

5 ☐ X_i følger en standard normal fordeling med middelværdi 0 og varians 3.

Spørgsmål V.3 (14)

Hvilken af de følgende R-kommandoer genererer 10 tilfældige tal fra en eksponentiel fordeling?

1 ☐ `replicate(10, rexp(1, 2))`

2 ☐ `pexp(seq(0.1, 1, length.out=10), 2)`

3 ☐ `qexp(seq(0.1, 1, 0.1), 2)`

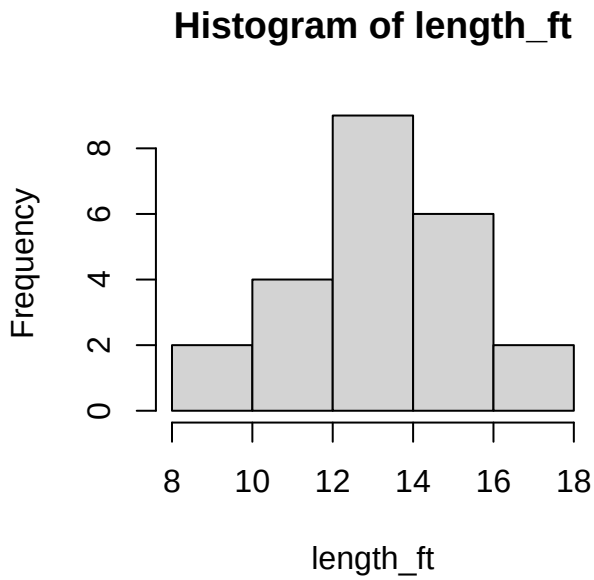
4 ☐ `rep(dexp(10, 2), 10)`

5 ☐ Inten af ovenstående. Den eksponentielle fordeling kræver en parameter til, som mangler i alle ovenstående svar.

Fortsæt på side 15

Opgave VI

Jesus Rivas, en herpetolog, forsker i grønne anacondas. Disse slanger, som er nogle af de største i verden, kan vokse op til 25 fod i længden. De har været kendt for at sluge levende geder og også mennesker. Jesus Rivas og medforskere vandrer barfodet på lavt vand i Llanos-græslandene, der deles af Venezuela og Colombia, i løbet af den tørre sæson. Når de føler en slange med fødderne, griber de den og holder den med hjælp af en anden person. Efter at have dæmpet slangen med en sok og tape, måler de slangens længde. 23 grønne anacondas blev fanget, og deres længde blev målt i fod. Dataene er gemt i `length_ft`. Du kan se det tilsvarende histogram af dataene herunder.



Spørgsmål VI.1 (15)

Hvilket af følgende er det korrekte 99% konfidensinterval for median-længden af anacondaer, hvis man antager at parametric bootstrapping blev brugt til at estimere intervallet?

```
median_ft <- median(length_ft)
mean_ft <- mean(length_ft)
sd_ft <- sd(length_ft)
n <- length(length_ft)
k <- 10000

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 12.02935 14.59873
```

```
sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
```

```r
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.94304 14.68206

sim_samples <- replicate(k, rchisq(n, mean_ft))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 10.75972 16.24469

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##     0.5%    99.5%
## 11.64546 15.04535

sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft^2))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))

##      0.5%     99.5%
##  9.121213 17.500782
```

1 ☐ [12.03, 14.60]

2 ☐ [11.94, 14.68]

3 ☐ [10.76, 16.24]

4 ☐ [11.65, 15.05]

5 ☐ [9.12, 17.50]

Spørgsmål VI.2 (16)

Hvilket af følgende er det korrekte 99% konfidensinterval for median-længden af anacondaer, hvis man antager at non-parametric bootstrapping blev brugt til at estimere intervallet?

```r
median_ft <- median(length_ft)
mean_ft <- mean(length_ft)
sd_ft <- sd(length_ft)
```

```r
n <- length(length_ft)
k <- 10000

sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 11.93076 15.22501
```

```r
sim_samples <- replicate(k, rnorm(n, mean_ft, sd_ft))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.005, 0.995))
```

```
##     0.5%    99.5%
## 11.61621 14.97613
```

```r
sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, mean)
quantile(sim_medians, c(0.01, 0.99))
```

```
##       1%      99%
## 12.08800 14.46791
```

```r
sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.01, 0.99))
```

```
##       1%      99%
## 12.48738 15.03513
```

```r
sim_samples <- replicate(k, sample(length_ft, n, replace = TRUE))
sim_medians <- apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 12.82957 14.46058
```

1 ☐ [11.93, 15.23]

2 ☐ [11.59, 15.05]

3 ☐ [12.13, 14.50]

4 ☐ [12.49, 15.04]

5 ☐ [12.83, 14.46]

Fortsæt på side 18

Opgave VII

Spørgsmål VII.1 (17)

Du har samlet rav med en ven, og I fandt i alt 20 stykker. I havde aftalt på forhånd, at dele dem ved tilfældigt at trække 10 stykker hver. Tre af stykkerne er meget attraktive. Hvad er sandsynligheden for, at du får alle tre attraktive stykker?

1 ☐ 0.0877%

2 ☐ 0.877%

3 ☐ 10.5%

4 ☐ 13.0%

5 ☐ 24.0%

Spørgsmål VII.2 (18)

Lad X repræsentere vægten i gram af et nyt stykke rav, som du finder på dit foretrukne samlested. Fra erfaring ved du, at når du finder et stykke rav der, så følger dets vægt en log-normal distribution, således at $X \sim LN(1, 0.7^2)$.

Hvad er gennemsnitsvægten μ_X af ravstykker på din favorit placering i henhold til denne model?

1 ☐ 2.01 g

2 ☐ 2.72 g

3 ☐ 3.47 g

4 ☐ 5.93 g

5 ☐ 9.21 g

Spørgsmål VII.3 (19)

Baseret på oplysningerne i det sidste spørgsmål: Hvis du finder 20 stykker rav på dit foretrukne samlested, hvad er sandsynligheden for, at mindst 3 af dem vejer over 10 gram?

1 ☐ 0.31%

2 ☐ 2.36%

3 ☐ 3.14%

4 ☐ 4.24%

5 ☐ 12.31%

Fortsæt på side 21

Opgave VIII

Lad den stokastiske variabel X_i repræsentere den i'te observation i en stikprøve med n observationer fra en population, der er uniform fordelt mellem α og β. Observationer trækkes tilfældigt og dermed uafhængigt af hinanden. Så $X_i \sim U(\alpha, \beta)$ og i.i.d.

Spørgsmål VIII.1 (20)

Stikprøvegennemsnittet er

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Hvad er fordelingen af $\bar{X}$ når n går mod uendeligt?

1 ☐ $N(0, 1^2)$

2 ☐ $U(\alpha, \beta)$

3 ☐ t-fordeling med $n - 1$ frihedsgrader

4 ☐ $N(\frac{\alpha+\beta}{2}, \frac{(\beta-\alpha)^2}{12n})$

5 ☐ $U(\alpha^n, \beta^n)$

Spørgsmål VIII.2 (21)

Definer $Y_i = 2 + \frac{1}{10} X_i$. Hvilket af følgende udsagn er korrekt?

1 ☐ $\mathrm{E}(Y_i) = \frac{1}{10} \mathrm{E}(X_i)$

2 ☐ $\mathrm{E}(Y_i) = \frac{1}{100} \mathrm{E}(X_i)$

3 ☐ $\mathrm{V}(Y_i) = \frac{1}{10} \mathrm{V}(X_i)$

4 ☐ $\mathrm{V}(Y_i) = \frac{1}{100} \mathrm{V}(X_i)$

5 ☐ $Y_i \sim U(\alpha, \beta)$

Fortsæt på side 22

Opgave IX

I elsystemer er regulerkraft den generation eller belastning, som kan øges eller reduceres hurtigt for at stabilisere spændingen på nettet. Regulerkraft handles ofte på et marked som det hollandske aFRR-marked, hvor bud afregnes i 15 minutters intervaller. Hvis man deltager på et sådan marked, er det vigtigt at vide hvor meget energi der aktiveres.

Først analyseres det aktiverede opreguleringsvolume, som er, hvor meget energi der i alt var aktiveret til øget generation pr. dag. De gennemsnitlige daglige værdier i MWh i tre vintermåneder læses ind i vektoren `xwinter` og følgende analyse udføres

```
t.test(xwinter)

##
##  One Sample t-test
##
## data:  xwinter
## t = 14, df = 89, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.346 12.341
## sample estimates:
## mean of x
##     10.84
```

Spørgsmål IX.1 (22)

Lad μ_{winter} være det gennemsnitlige opreguleringsvolume på vinterdage. Hvis man anvender et signifikansniveau $\alpha = 0.05$, hvad bliver da konklusionen for følgende nulhypotese (både konklusion og argument skal være korrekte)?

$$H_0 : \mu_{\text{winter}} = 10$$

1 ☐ Nulhypotesen afvises, da p-værdien er mindre end $2 \cdot 10^{-16}$ hvilket er mindre end 5%

2 ☐ Nulhypotesen accepteres, da p-værdien er mindre end $2 \cdot 10^{-16}$ hvilket er mindre end 5%

3 ☐ Nulhypotesen afvises, da p-værdien er mindre end $2 \cdot 10^{-16}$ hvilket er større end 5%

4 ☐ Nulhypotesen accepteres, da p-værdien er mindre end $2 \cdot 10^{-16}$ hvilket er større end 5%

5 ☐ Nulhypotesen accepteres, da 10 er indeholdt i 95% konfidensintervallet

Spørgsmål IX.2 (23)

Hvad er 99% konfidensintervallet for μ_{winter}?

1 ☐ $[7.77, 13.91]$

2 ☐ $[8.01, 12.10]$

3 ☐ $[8.28, 13.41]$

4 ☐ $[8.86, 12.82]$

5 ☐ $[9.35, 12.34]$

Spørgsmål IX.3 (24)

Hvad er antallet af observationer i `xwinter`?

1 ☐ 88

2 ☐ 89

3 ☐ 90

4 ☐ 91

5 ☐ 92

Spørgsmål IX.4 (25)

For at finde ud af, om der er forskel mellem vinter og sommer, indlæses de daglige gennemsnit af opreguleringsvolume for sommermånederne det samme år i `xsummer`.

Baseret på de givne data her i opgaven, hvilken af følgende tests er bedst egnet til at konkludere, om der er en signifikant forskel mellem den daglige middelværdi af opreguleringsvolume om vinteren og om sommeren?

1 ☐ En two-sample t-test

2 ☐ En parret two-sample t-test

3 ☐ En tovejs ANOVA test

4 ☐ En test for hældningskoefficienten i en lineær regressionsmodel

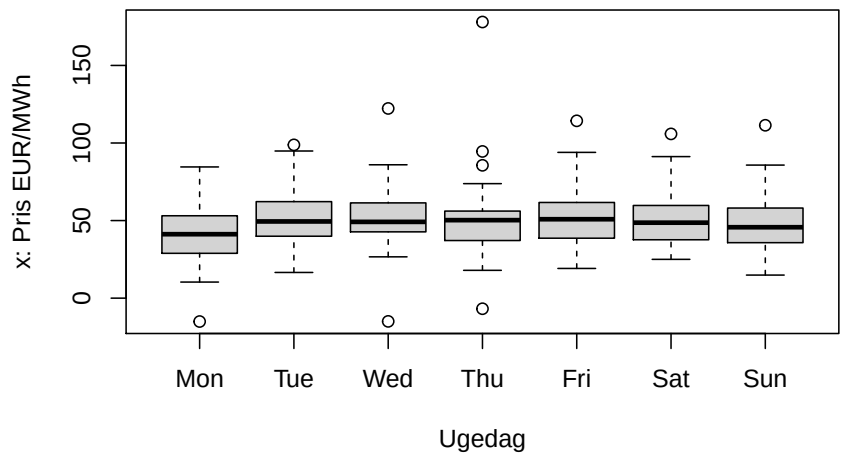
5 ☐ En χ^2-test

Fortsæt på side 24

Opgave X

Denne opgave handler om det hollandske marked for reguleringskraft, som beskrevet i forrige opgave. For udbydere af reguleringskraft er det vigtigt at undersøge priserne for salg og køb på markedet. Et års daglige gennemsnitspriser på nedregulering læses ind i `x`. 364 observationer (dage) er inkluderet i data.

For at se, om der er forskelle mellem ugedagene, genereres der boxplots for hver ugedag (bemærk, at priserne er angivet pr. energienhed, denne detalje betyder ikke noget for opgaven):



En en-vejs ANOVA er udført og resultat er:

```
anova(lm(x ~ weekday))

## Analysis of Variance Table
##
## Response: x
##            Df Sum Sq Mean Sq F value  Pr(>F)
## weekday     6   4934  822.42  2.0969 0.05296 .
## Residuals 357 140016  392.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål X.1 (26)

Ved et signifikansniveau på 5%, hvad er da den kritiske værdi for F-testen for ens middelværdi på ugedagene?

1 ☐ 1.549

2 ☐ 1.791

3 ☐ 1.943

4 ☐ 2.124

5 ☐ 2.444

Spørgsmål X.2 (27)

Under antagelse af, at alle forudsætninger for den anvendte model er opfyldt, hvad er da estimatet af variansen af den daglige nedreguleringspris på fredage (både værdien og forklaringen skal være korrekt)?

1 ☐ $\hat{\sigma}^2 = 392.2$, da variansestimatet er sammenvejet (pooled) og således ens for alle ugedage

2 ☐ $\hat{\sigma}^2 = \frac{140016}{4934} = 28.38$, da variansestimatet er sammenvejet (pooled) og således ens for alle ugedage

3 ☐ $\hat{\sigma}^2 = \frac{140016}{7} = 20002$, da variansestimatet skal estimeres individuelt for ugedagene, og derfor tilpasset med antal frihedsgrader for `weekdays`

4 ☐ $\hat{\sigma}^2 = \frac{140016}{6} = 23336$, da variansestimatet skal estimeres individuelt for ugedagene, og derfor tilpasset med antal frihedsgrader for `weekdays`

5 ☐ Dette kan ikke udregnes med den givne information

Spørgsmål X.3 (28)

Hvor stor en andel af variansen er forklaret af modellen?

1 ☐ 0.57%

2 ☐ 3.4%

3 ☐ 18.4%

4 ☐ 32.3%

5 ☐ 96.6%

Fortsæt på side 26

Spørgsmål X.4 (29)

Nu tages ugenummeret `week` med som en forklarende variabel og en to-vejs ANOVA udføres:

```
anova(lm(x ~ weekday + week))

## Analysis of Variance Table
##
## Response: x
##            Df Sum Sq Mean Sq F value Pr(>F)
## weekday     6   4934     822    2.97 0.0079 **
## week       51  55218    1083    3.91  6e-14 ***
## Residuals 306  84798     277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ved sammenligning af dette resultat med resultatet af en-vejs ANOVA analysen, ses det, at p-værdien for `weekday` er faldet meget. Hvilket af følgende udsagn er den rigtige forklaring på dette?

1 ☐ Variansen forklaret ved gruppering på ugedage ($SS(weekday)$) stiger, således at p-værdien for effekten af ugedage falder

2 ☐ Frihedsgraderne for `Residuals` falder, hvilket fører til faldet i p-værdien for effekten ugedage

3 ☐ Variansen forklaret ved gruppering på ugedage ($SS(weekday)$) falder, således at p-værdien for effekten af ugedage falder

4 ☐ Residualernes kvadratafvigelsessum (SSE) falder signifikant, hvilket fører til faldet i p-værdien for effekten ugedage

5 ☐ Der må være en signifikant sammenhæng mellem ugedagenes middelværdier og ugernes middelværdier, hvilket fører til faldet i p-værdi for effekten af ugedage

Spørgsmål X.5 (30)

Vi er nu interesserede i at udføre en post-hoc analyse med ANOVA modellen fra forrige spørgsmål. Følgende blev kørt i R:

```
tapply(x, weekday, mean)

##  Mon  Tue  Wed  Thu  Fri  Sat  Sun
## 41.4 51.3 52.0 50.9 52.8 51.7 48.1
```

```
tapply(x, weekday, sd)

##  Mon  Tue  Wed  Thu  Fri  Sat  Sun
## 19.9 18.3 19.5 24.4 20.2 17.4 18.0
```

Hvilket af følgende R kald giver det korrekte enkelt-forudplanlagte 95% konfidensinterval for forskellen i middel af den daglige pris på lørdage og på søndage?

1 ☐ `t.test(x[weekday=="Sat"], x[weekday=="Sun"], conf.level=0.9976)`

2 ☐ `t.test(x[weekday=="Sat"], x[weekday=="Sun"])`

3 ☐ `51.7 - 48.1 + c(-1,1) * qt(0.975, 306) * sqrt(2 * 277 * 1/52)`

4 ☐ `51.7 - 48.1 + c(-1,1) * qt(0.9988, 52) * sqrt(2 * 277 * 1/52)`

5 ☐ `51.7 - 48.1 + c(-1,1) * qt(0.9988, 306) * sqrt(17.4^2/52 + 18.0^2/52)`

Fortsæt på side 28

SÆTTET ER SLUT. God jul!