Danmarks Tekniske Universitet *Side 1 af 27 sider.*

Skriftlig prøve: 21. Juni 2018

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

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

_____	_____	_____
(studienummer)	(underskrift)	(bord nr)

Opgavesættet består af 30 spørgsmål af "multiple choice" typen fordelt på 15 opgaver. Besvarelserne af "multiple choice" spørgsmålene anføres i det i CampusNet uploadede svarark (på 6 seperate sider), med numrene på de svarmuligheder, du mener er de korrekte.

Der gives 5 point for et korrekt "multiple choice" svar og -1 point for et ukorrekt svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller anden type svar angives, tæller det ikke med i besvarelsen. Envidere, hvis mere end et svar angives, hvilket faktisk er teknisk muligt i online-systemet, så tæller det ikke med (dvs. spørgsmålet giver "0 point"). Det antal point der kræves, for at et sæt anses for tilfredstillende besvaret, afgøres endeligt ved censureringen.

Den endelige besvarelse af opgaverne gøres ved at udfylde og online-aflevere svararket via CampusNet. Skemaet her er KUN et nød-alternativ til dette. Husk at angive dit studienummer på din besvarelse, hvis du afleverer skemaet.

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

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

Opgave	XI.3	XI.4	XI.5	XI.6	XII.1	XIII.1	XIII.2	XIV.1	XV.1	XV.2
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Sættet består af 27 sider.

Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at ideen med opgaverne er, at der er ét og kun ét rigtigt svar på hvert spørgsmål. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde.*

Opgave I

En terning kastes på et havebord med en vis afstand mellem plankerne. Sandsynligheden for at terningen ender på en kant mellem to planker er 0.2. Hvis terningen ikke ender mellem to planker, så har den lige stor sandsynlighed for at lande på enhver af de 6 sider. Hvis terningen ender på kanten mellem 2 planker så tilskrives den værdien 0.

Spørgsmål I.1 (1)

Hvad er forventningværdien for fordelingen svarende til et terningkast?

1 ☐ 2.5

2 ☐ 2.8

3 ☐ 3

4 ☐ 3.3

5 ☐ 3.5

Fortsæt på side 3

Opgave II

10 individer er delt i to grupper med 5 individer i hver. Hvert individ har fået foretaget en celletælling, som en del af en medicinsk undersøgelse. Antallene er givet i tabellen herunder

Group 1	351	276	94	293	364
Group 2	494	403	159	329	492

Det kan i det følgende antages at variansen i de to grupper er ens, samt at data stammer fra normalfordelinger.

Spørgsmål II.1 (2)

Idet notens definition af fraktiler benyttes, hvad er så 1., 2. og 3. kvartil af data i gruppe 1?

1 ☐ (276, 94, 293)

2 ☐ (276, 351, 351)

3 ☐ (276, 293, 351)

4 ☐ (185, 293, 357.5)

5 ☐ (276, 351, 364)

Spørgsmål II.2 (3)

Hvis vi antager, at de to grupper er uafhængige, hvad er så p-værdien for en standard t-test for om der er forskel i tælletallene i de to grupper?

1 ☐ 0.44

2 ☐ 0.34

3 ☐ 0.14

4 ☐ 0.04

5 ☐ 0.24

Fortsæt på side 4

Spørgsmål II.3 (4)

Hvis nu de to grupper i stedet refererer til to forskellige metoder til at tælle celler på anvendt på 5 personer, således at søjler refererer til personer, hvad ville p-værdien for om, der er forskel mellem de to grupper, så være?

1 ☐ 0.299

2 ☐ 0.049

3 ☐ 0.199

4 ☐ 0.099

5 ☐ 0.009

Fortsæt på side 5

Opgave III

I forbindelse med de netop overståede overenskomstforhandlinger har der været et stort fokus på løn. Danmarks 98 kommuner er inddelt i fem regioner: Hovedstaden, Midtjylland, Nordjylland, Sjælland og Syddanmark. I det følgende analyseres lærernes månedsløn i 1.000 kr.

Nedenfor vises 2 analyser af data i R (`lldr` indeholder data for lærenes løn som beskrevet ovenfor):

```
lm.lldr <- lm(loen ~ Region, lldr)
summary(lm.lldr)

##
## Call:
## lm(formula = loen ~ Region, data = lldr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14456 -0.41405 -0.03269  0.49197  1.98813
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        42.6325     0.1530 278.603  < 2e-16 ***
## RegionMidtjylland  -0.6566     0.2432  -2.700  0.00825 **
## RegionNordjylland  -0.5423     0.2918  -1.859  0.06625 .
## RegionSjælland     -0.4222     0.2517  -1.677  0.09685 .
## RegionSyddanmark   -0.5352     0.2330  -2.297  0.02386 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.824 on 93 degrees of freedom
## Multiple R-squared:  0.0932,Adjusted R-squared:  0.0542
## F-statistic:  2.39 on 4 and 93 DF,  p-value: 0.0564

anova(lm.lldr)

## Analysis of Variance Table
##
## Response: loen
##           Df Sum Sq Mean Sq F value Pr(>F)
## Region     4  6.491 1.62270  2.3896 0.0564 .
## Residuals 93 63.152 0.67906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fortsæt på side 6

Spørgsmål III.1 (5)

Den estimerede model betegnes typisk som en

1 ☐ Simpel lineær regressionsmodel

2 ☐ To-vejs variansanalyse

3 ☐ T-test

4 ☐ En-vejs variansanalyse

5 ☐ Ingen af ovenstående

Spørgsmål III.2 (6)

Idet man benytter det sædvanlige signifikansniveau ($\alpha = 0.05$), bliver konklusionen på analysen at

1 ☐ der er er signifikant forskel på regionerne, da $2.39 > 1.96$

2 ☐ der ikke er signifikant forskel på regionerne, da $0.0564 > 0.05$

3 ☐ der er meget signifikant forskel på regionerne, da $0.00825 < 0.05$

4 ☐ der er signifikant forskel på regionerne, da $0.0564 > 0.05$

5 ☐ der mangler data fra én kommune, da $93 + 4 \neq 98$.

Fortsæt på side 7

Opgave IV

I nedenstående tabel er de gennemsnitlige lønninger for personer med en mellemlang videregående uddannelse angivet for en periode på 6 år og inddelt i de fem regioner.

Region	2011	2012	2013	2014	2015	2016
Hovedstaden	31.36	31.93	32.49	33.02	34.77	34.50
Sjælland	31.68	32.11	32.16	32.87	33.28	33.62
Syddanmark	30.84	31.36	31.40	31.91	32.57	32.76
Midtjylland	29.87	30.12	30.62	30.74	31.52	32.00
Nordjylland	29.28	29.63	29.82	30.10	30.42	30.65

Resultatet af den relevante analyse er:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Region       4   3778   944.6       A        C
year         5   1734   346.8       B        D
Residuals   20    242    12.1
```

(Enkelte tal er udeladt og erstattet af "A", "B", "C" og "D").

Spørgsmål IV.1 (7)

Den relevante teststørrelse for at undersøge effekten af år findes til

1 ☐ 1737 / 346.8 = 5.009

2 ☐ 346.8 / 12.1 = 28.661

3 ☐ 5 / 1734 = 0.003

4 ☐ 944.6 / 346.8 = 2.724

5 ☐ 1734 / 242 = 7.165

Fortsæt på side 8

Spørgsmål IV.2 (8)

Teststørrelsen for testet for forskelle mellem regioner skal sammenlignes med en

1 ☐ χ^2 fordeling med 20 frihedsgrader

2 ☐ t-fordeling med 4 frihedsgrader

3 ☐ F(4, 5) fordeling

4 ☐ χ^2 fordeling med 4 frihedsgrader

5 ☐ F(4, 20) fordeling

Fortsæt på side 9

Opgave V

Antag at der er lavet en en-stikprøve test med 15 observationer.

Spørgsmål V.1 (9)

Hvad er den kritiske værdi for en standard t-test på signifikansniveau $\alpha = 0.01$ for om den underliggende fordeling har middelværdi 0 mod alternativet at den er positiv?

1 ☐ 2.98=qt(0.995,df=14)

2 ☐ 2.62=qt(0.99,df=14)

3 ☐ 1.75=qt(0.95,df=15)

4 ☐ 2.60=qt(0.99,df=15)

5 ☐ 2.95=qt(0.995,df=15)

Fortsæt på side 10

Opgave VI

I en indeklimaundersøgelse ønskes forhold i bygninger undersøgt. Der er specielt fokus på oplevelsen, som bygningernes brugere har, og derfor gennemføres en række spørgeskemaundersøgelser. I det følgende skal udføres statistik på resultaterne af undersøgelsen.

Spørgsmål VI.1 (10)

I en kontorbygning, hvor der har været udført en renovering, blev der gennemført en undersøgelse før og en efter renoveringen. Der blev bl.a. stillet følgende spørgsmål: "Har du følt dig tør i halsen på kontoret i løbet af den sidste uge?"

	Før	Efter
Hele tiden	34	22
En del af tiden	39	43
På intet tidspunkt	42	34

Følgende R kode er kørt:

```
prop.test(x=c(34,22), n=c(34+39+42,22+43+34), correct=FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(34, 22) out of c(34 + 39 + 42, 22 + 43 + 34)
## X-squared = 1.4847, df = 1, p-value = 0.223
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04345776  0.19031766
## sample estimates:
##    prop 1    prop 2
## 0.2956522 0.2222222

chisq.test(matrix(c(34, 22, 39, 43, 42, 34), ncol = 2, byrow = TRUE),
           correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  matrix(c(34, 22, 39, 43, 42, 34), ncol = 2, byrow = TRUE)
## X-squared = 2.426, df = 2, p-value = 0.2973
```

Fortsæt på side 11

Det ønskes nu undersøgt, om hele fordelingen af svar på spørgsmålet er forskellig før renoveringen i forhold til efter renoveringen. Hvad er konklusionen på signifikansniveau $\alpha = 0.05$ (både konklusion og argument skal være rigtigt)?

1 ☐ Nulhypotesen om at der ikke er forskel på fordelingen før og efter accepteres, da p-værdien for den relevante test er 0.223

2 ☐ Nulhypotesen om at der ikke er forskel på fordelingen før og efter afvises, da p-værdien for den relevante test er 0.223

3 ☐ Nulhypotesen om at der ikke er forskel på fordelingen før og efter accepteres, da p-værdien for den relevante test er 0.2973

4 ☐ Nulhypotesen om at der ikke er forskel på fordelingen før og efter afvises, da p-værdien for den relevante test er 0.2973

5 ☐ Ingen af ovenstående svarmuligheder er korrekte.

Spørgsmål VI.2 (11)

I et multirumskontor ønskes indeklimaet på de enkelte kontorpladser undersøgt i forhold til hinanden. Derfor spørges de ansatte løbende over tid, via en app, om de syntes indeklimaet er godt. F.eks. får de spørgsmålet: "Har der været for varmt på din kontorplads i dag?". I løbet af en uge er følgende svar til dette spørgsmål indsamlet for kontorpladserne tæt på vinduet:

Ja	28
Nej	17

Hvilket af følgende svar er det korrekt beregnede 99% konfidensinterval for andelen af medarbejdere på vinduespladserne, som har haft det for varmt?

1 ☐ $\frac{28}{45} \pm 1.96\sqrt{\frac{476}{2025}}$

2 ☐ $\frac{28}{45} \pm 2.58\sqrt{\frac{476}{91125}}$

3 ☐ $\frac{28}{45} \pm 2.58\sqrt{\frac{476}{2025}}$

4 ☐ $\frac{28}{45} \pm 1.96\sqrt{\frac{476}{91125}}$

5 ☐ $\frac{28}{45} \pm 2.58\sqrt{\frac{784}{2025}}$

Fortsæt på side 12

Spørgsmål VI.3 (12)

Undersøgelsen fra forrige spørgsmål køres løbende, og der ønskes en forventet bredde af konfidensintervallet på 20% på signifikansniveau $\alpha = 0.05$.

Det antal svar der mindst skal indsamles afhænger af scenariet (dvs. den antagede sandsynlighed, p, for positive svar). Hvor mange færre svar skal der indsamles i et scenarie, hvor det antages at $p = 0.25$ i forhold til det mest konservative scenarie, dvs. med antagelsen $p = 0.5$?

1 ☐ 6 færre svar

2 ☐ 24 færre svar

3 ☐ 35 færre svar

4 ☐ 73 færre svar

5 ☐ 100 færre svar

Fortsæt på side 13

Opgave VII

I en undersøgelse ønskes det belyst, om der er en sammenhæng mellem kunders opfattelse af, hvor grønne deres indkøbsvaner er, i forhold til deres faktiske indkøb, målt på hvor mange økologiske varer de køber. Kunderne har via en supermarkeds-app svaret på spørgsmålet: "Jeg tænker meget på miljøet, når jeg køber ind" og supermarkedet har registreret kundernes indkøb og inddelt dem i 3 kategorier efter, hvor meget økologisk de har købt.

Økologisk indkøb	*Jeg tænker på miljøet når jeg køber ind*		
	Enig	Hverken enig eller uenig	Uenig
Meget	21.5%	6.8%	5.1%
Middel	17.5%	13%	6.8%
Lavt	9.6%	11.9%	7.9%

Der blev lavet ialt indsamlet svar fra 177 kunder (at cellerne ikke summerer til præcis 100% skyldes almindelig afrunding).

Spørgsmål VII.1 (13)

Under nulhypotesen om uafhængighed i kundens opfattelse og handling, hvad er da det forventede antal observationer for kunder, med økologisk indkøb "Meget" og som angiver "Enig" i at tænke på miljøet, når de køber ind (dvs. øverste venstre celle)?

1 ☐ $(0.068 + 0.13 + 0.119) \cdot 177 \cdot (0.175 + 0.13 + 0.068) \cdot 177/177 = 20.9$

2 ☐ $(0.215 + 0.068 + 0.051) \cdot 177 \cdot (0.215 + 0.175 + 0.096) \cdot 177/177 = 28.7$

3 ☐ $(0.215) \cdot 177 \cdot (0.175 + 0.13 + 0.068) \cdot 177/177 = 14.2$

4 ☐ $(0.215 + 0.068 + 0.051) \cdot 177 = 59.1$

5 ☐ $(0.175 + 0.13 + 0.068) \cdot 177 = 66$

Spørgsmål VII.2 (14)

Antallet af frihedsgrader (df) og den kritiske værdi (q_{crit}) ved det relevante test på signifikansniveau $\alpha = 0.05$ bliver:

1 ☐ $df = 177$ og $q_{\text{crit}} = 1.96$

2 ☐ $df = 168$ og $q_{\text{crit}} = 3.84$

3 ☐ $df = 2$ og $q_{\text{crit}} = 5.99$

4 ☐ $df = 4$ og $q_{\text{crit}} = 9.49$

5 ☐ $df = 3$ og $q_{\text{crit}} = 7.82$

Fortsæt på side 14

Opgave VIII

I et lotteri er 7 ud af 36 vindertal. En lodseddel består af 7 forskellige tilfældigt udvalgte (af de 36) tal. Hvis en lodseddel indeholder mindst 4 vindertal udløses en præmie.

Spørgsmål VIII.1 (15)

Hvad er sandsynligheden for at vinde en præmie, hvis man køber en lodseddel?

1 ☐ 0.00062%

2 ☐ 0.031%

3 ☐ 0.54%

4 ☐ 1.64%

5 ☐ 2.5%

Fortsæt på side 15

Opgave IX

Herunder ses observationer af tilbagevendelsestiden for brystkræft (målt i den naturlige logaritme til måneder, dvs. log(måneder)).

Patient	1	2	3	4	5
log(måneder)	3.3	4.34	4.34	3.58	2.30

Det kan antages at observationerne i tabellen er uafhængige og normalfordelte ($N(\mu, \sigma^2)$), samt at gennemsnit og empirisk varians er hhv. $\hat{\mu} = 3.57$ og $s^2 = 0.72$.

Spørgsmål IX.1 (16)

Baseret på ovenstående estimater, hvad er da den forventede tilbagevendelsestid i måneder for brystkæft?

1 ☐ 73

2 ☐ 54

3 ☐ 51

4 ☐ 36

5 ☐ 25

Spørgsmål IX.2 (17)

Hvad er et 95% konfidensinterval for variansparameteren (σ^2)?

1 ☐ $[0.32; 7.43]$

2 ☐ $[0.26; 5.95]$

3 ☐ $[3.07; 4.06]$

4 ☐ $[0.31; 4.05]$

5 ☐ $[2.52; 4.62]$

Fortsæt på side 16

Opgave X

To grupper af observationer af skrivehastighed (målt i ord per minut) er givet i tabellen herunder

Gruppe 1	35	50	55	60	65	60	70	55	45	55	60	45	65	55	50	60
Gruppe 2	55	60	75	65	60	70	75	70	65	72	73	65	80	50	55	70

Det oplyses desuden at stikprøvevariansen i de to grupper er hhv. $s_1^2 = 78.23$ og $s_1^2 = 70.87$, samt at antallet af observationer i hver gruppe er $n_1 = n_2 = 16$.

Spørgsmål X.1 (18)

Hvad er antallet af frihedsgrader for en Welch t-test for om de to grupper har samme skrivehastighed?

1 ☐ 16

2 ☐ 18.9

3 ☐ 23.2

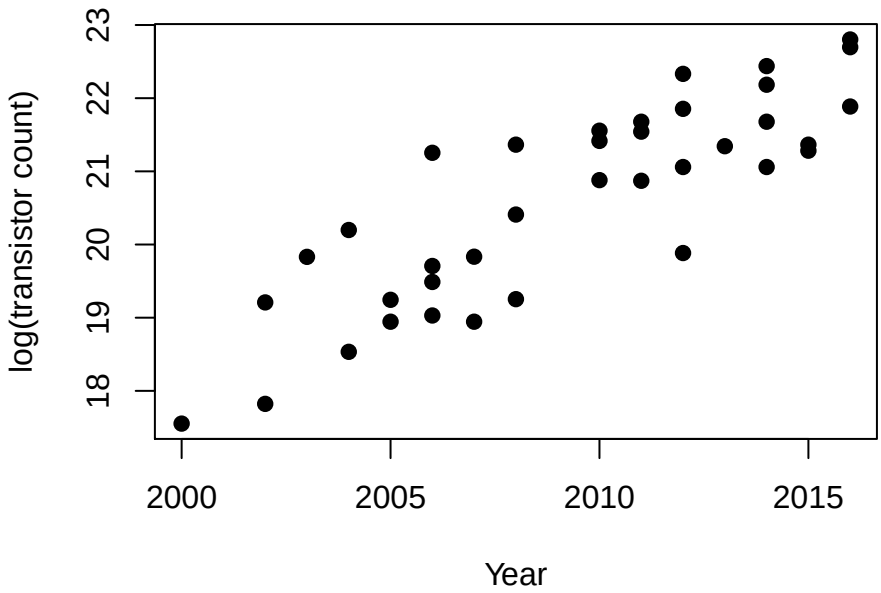
4 ☐ 29.9

5 ☐ 32

Fortsæt på side 17

Opgave XI

Antallet af transistorer ("transistor count") på et integreret kredsløb er det mest almindelige mål for kompleksiteten af integrerede kredsløb. Figuren herunder viser den naturlige logaritme af "transistor count", som funktion af årstal siden år 2000 for et udvalg af transistorer, der kom på markedet de enkelte år.



Figuren antyder at en log-linear model for antallet af "counts" kunne være passende. For at undersøge hypetesen har man kørt følgende R-code (hvor variablen `year`=Årstal-2000):

```
summary(fit <- lm(log(count) ~ year))

##
## Call:
## lm(formula = log(count) ~ year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4239 -0.6099 -0.1532  0.5468  1.5184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.16313    0.27627  65.743  < 2e-16 ***
## year         0.26206    0.02676   9.792 1.47e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7212 on 35 degrees of freedom
## Multiple R-squared:  0.7326,Adjusted R-squared:  0.7249
## F-statistic: 95.88 on 1 and 35 DF,  p-value: 1.465e-11
```

Spørgsmål XI.1 (19)

Hvis signifikansniveau $\alpha = 0.05$ benyttes, hvilket af følgende udsagn omkring modellen er så korrekt (både konklusion og argument skal være korrekt)?

1 ☐ Afskæringen med y-aksen er ikke signifikant forskellig fra 0 da $0.28 > 0.05$

2 ☐ Afskæringen med y-aksen er signifikant forskellig fra 0 da $0.027 < 0.05$

3 ☐ Afskæringen med y-aksen er signifikant forskellig fra 0 da $2 \cdot 10^{-16} < 0.05$

4 ☐ Afskæringen med y-aksen er ikke signifikant forskellig fra 0 da $2 \cdot 10^{-16} < 0.05$

5 ☐ Afskæringen med y-aksen er signifikant forskellig fra 0 da $1.47 \cdot 10^{-11} < 0.05$

Spørgsmål XI.2 (20)

Et ofte diskuteret emne er den relative forøgelse af transistor count pr. år, dvs. $\gamma = \frac{\hat{Y}_{year+1}}{\hat{Y}_{year}}$, hvor $\hat{Y}_{year}$ er det forventede count i year. Hvad er et 95% konfidensinterval for konstanten γ?

1 ☐ $[0.21; 0.32]$

2 ☐ $[18.37; 18.47]$

3 ☐ $[1.23; 1.37]$

4 ☐ $[17.31; 19.54]$

5 ☐ Det kan man ikke svare på med de givne oplysninger

Til brug for de videre udregninger har man kørt følgende R-kode

```
c(mean(year), var(year))
```

```
## [1]  9.324324 20.169670
```

Fortsæt på side 19

Spørgsmål XI.3 (21)

Hvad er, baseret på modellen, det sædvanlige 95% konfidensinterval for $\log(count)$ i år 2020?

1 ☐ $[21.94; 24.87]$

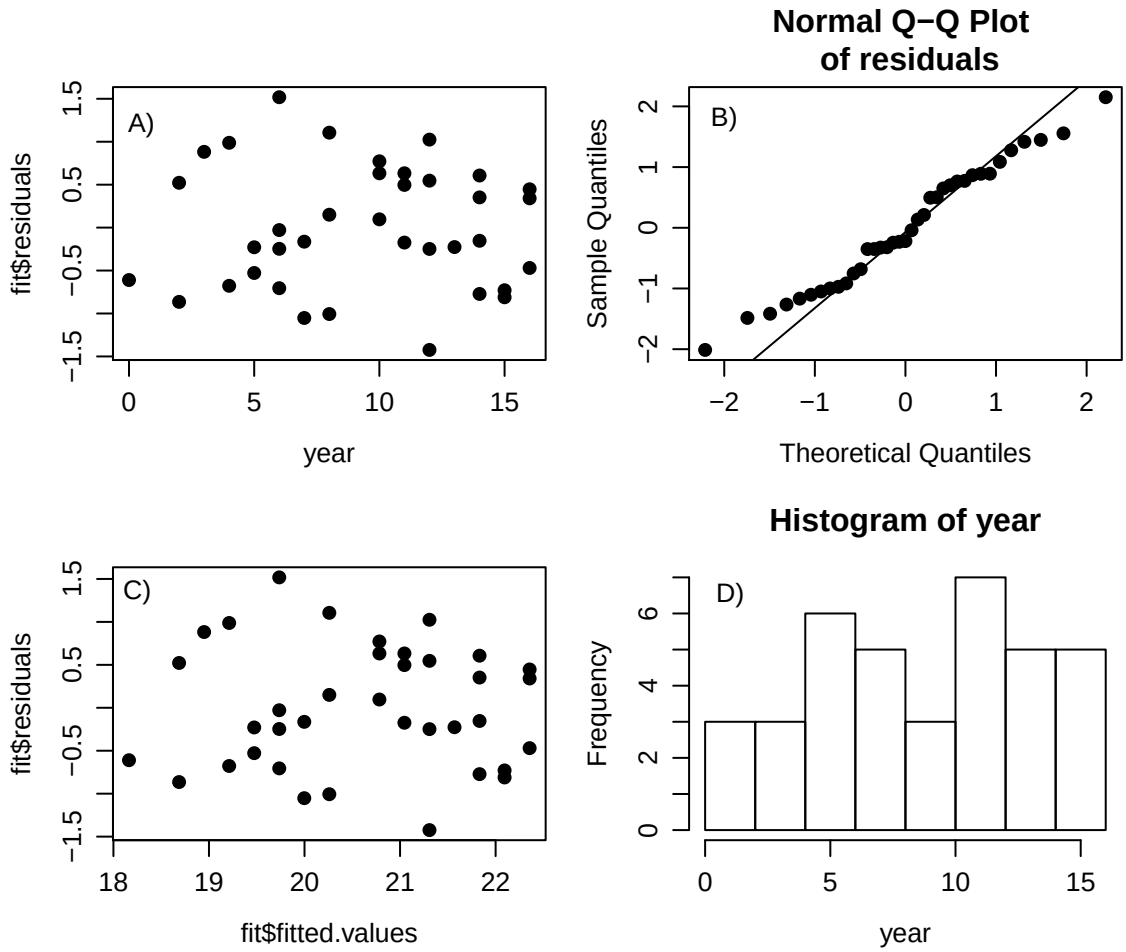
2 ☐ $[438.27; 656.78]$

3 ☐ $[19.62; 27.19]$

4 ☐ $[22.78; 24.03]$

5 ☐ $[529.06; 565.99]$

Til brug for undersøgelse af modellens forudsætninger har man lavet nedenstående figur.



Fortsæt på side 20

Spørgsmål XI.4 (22)

Hvilket af følgende udsagn om modellen kan, baseret på figuren, være korrekt (både konklusion, figurhenvisning, og argument skal være korrekt)?

1 ☐ Da residualerne som funktion af `year` ikke ligger på ret linie (Plot A), er modellens forudsætninger klart ikke opfyldt.

2 ☐ Da variablen `year` ikke følger en normalfordeling (Plot D), er modellens forudsætninger ikke opfyldt.

3 ☐ Antagelsen om uafhængighed er klart ikke opfyldt (Plot B)

4 ☐ Da variablen `year` tilnærmelsesvis følger en normalfordeling (Plot D), er modellens forudsætninger opfyldt.

5 ☐ Antagelsen om varians homogenitet ser ud til at være opfyldt (Plot C)

Som en første undersøgelse af om hældningen ændrer sig over tid har man kørt følgende model i R (hvor `year2`=`year`^{2})

```
fit <- lm(log(count)~year+year2)
```

Modellen kan skrives i matrix-vektor notation som

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

hvor $\boldsymbol{X}$ er den såkaldt design matrix.

Spørgsmål XI.5 (23)

Hvilket af følgende udsagn, om designmatricen svarende til modellen i R overnfor, er korrekt?

1 ☐ Første kolonne indeholder log-counts

2 ☐ Anden kolonne består af vektoren `year2`

3 ☐ Første kolonne består af vektoren `year`

4 ☐ Første kolonne er en vektor af et-taller

5 ☐ Designmatricen har 3 rækker

Fortsæt på side 21

Til brug for undersøgelse af parametrene i modellen har man nu kørt følgende R-kode (hvor `X` er designmatricen)

```
solve(t(X) %*% X)

##               [,1]         [,2]          [,3]
## [1,]  0.405562066 -0.091123952  4.420991e-03
## [2,] -0.091123952  0.025056573 -1.337289e-03
## [3,]  0.004420991 -0.001337289  7.552324e-05
```

Spørgsmål XI.6 (24)

Hvis vi benævner modellens middelværdiparametre $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$, hvad er så estimatet for korrelationen mellem $\hat{\beta}_1$ og $\hat{\beta}_2$?

1 ☐ -0.0911

2 ☐ -0.904

3 ☐ -0.00133

4 ☐ -0.972

5 ☐ 0.00442

Fortsæt på side 22

Opgave XII

Lad X og Y være stokastiske variable med varianser hhv. σ_X^2 og σ_Y^2. Lad desuden korrelationen mellem X og Y være ρ. Konstruer en stokastisk variabel $Z = X - aY$ hvor $a \in \mathbb{R}$.

Spørgsmål XII.1 (25)

Hvilken værdi skal a være for at Z og Y er ukorreleret?

1 ☐ ρ

2 ☐ 0

3 ☐ $\rho\sigma_x\sigma_y$

4 ☐ $\rho\sigma_x/\sigma_y$

5 ☐ $\rho\sigma_y^2$

Fortsæt på side 23

Opgave XIII

Som en del af et biologisk forsøg har man undersøgt en bestemt type bakteries vækst i en petriskål. Petriskålen blev inddelt i 20 lige store felter, og antallet af bakteriekolonier i hvert felt blev talt. Dette gav anledning til følgende 20 observationer, som er indlæst i R i vektoren `bacteria`:

```
bacteria

##  [1] 7 6 4 6 2 2 7 1 2 5 7 6 3 3 6 6 2 6 4 5
```

Observationerne antages at være uafhængige udfald fra en Poissonfordeling med middelværdi λ.

Spørgsmål XIII.1 (26)

Man beslutter indledningsvis at analysere data uden at bruge Poissonfordelingsantagelsen. Hvilken af følgende stykker R-kode beregner et 95% konfidensinterval for standardafvigelsen for fordelingen af antallet af bakteriekolonier i hvert felt ved hjælp af ikke-parametrisk bootstrap?

1 ☐
```
counts <- replicate(10000, sample(bacteria, replace = TRUE))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.05, 0.95))
```

2 ☐
```
counts <- replicate(10000, rpois(20, mean(bacteria)))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.025, 0.975))
```

3 ☐
```
counts <- replicate(10000, rpois(20, mean(bacteria)))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.05, 0.95))
```

4 ☐
```
counts <- replicate(10000, sample(bacteria, replace = TRUE))
var_count <- apply(counts, 2, var)
quantile(sqrt(var_count), c(0.025, 0.975))
```

5 ☐
```
counts <- replicate(10000, rnorm(20, mean(bacteria), sd(bacteria)))
sd_count <- apply(counts, 2, sd)
quantile(sd_count, c(0.025, 0.975))
```

Spørgsmål XIII.2 (27)

Hvad beregnes i nedenstående R-kode?

```
counts <- replicate(10000, rpois(20, mean(bacteria)))
y <- apply(counts, 2, sd)/apply(counts, 2, mean)
quantile(y, c(1-0.95, 0.95))
```

1 ☐ Et 95% konfidensinterval for variationskoefficienten for fordelingen af antallet af bakteriekolonier i hvert felt, beregnet ved hjælp af parametrisk bootstrap.

2 ☐ Forholdet mellem 90% konfidensintervallerne for henholdsvis standardafvigelsen og middelværdien af fordelingen af antallet af bakteriekolonier i hvert felt, beregnet ved hjælp af parametrisk bootstrap.

3 ☐ Et 90% konfidensinterval for variationskoefficienten for fordelingen af antallet af bakteriekolonier i hvert felt, beregnet ved hjælp af parametrisk bootstrap.

4 ☐ 90% konfidensintervaller for standardafvigelsen og middelværdien af fordelingen af antallet af bakteriekolonier i hvert felt, beregnet ved hjælp af parametrisk bootstrap.

5 ☐ Et 90% konfidensinterval for variationskoefficienten for fordelingen af antallet af bakteriekolonier i hvert felt, beregnet ved hjælp af ikke-parametrisk bootstrap.

Fortsæt på side 25

Opgave XIV

Som en del af et biologisk forsøg har man undersøgt to forskellige typer bakteriers vækst i en petriskål. Petriskålen blev inddelt i 15 lige store felter, og antallet af bakteriekolonier i hvert felt blev talt. Dette gav anledning til følgende 15 observationer for hver af de to bakterietyper, som er indlæst i R i vektorerne `bact1` og `bact2`:

```
bact1

##  [1] 8 5 5 4 7 7 1 4 5 4 6 0 1 4 5

bact2

##  [1]  5  5  7 10  8  7  5 10  7  4  9  6  5  3  4
```

Observationerne antages at være uafhængige. Antal bakteriekolonier i hvert felt antages at være Poissonfordelt med ukendt middelværdi λ_i for bakterietype i (i er her 1 eller 2).

Spørgsmål XIV.1 (28)

Angiv et stykke R-kode som beregner et 99% parametrisk bootstrap konfidensinterval for forskellen i middelantal bakteriekolonier per felt mellem de to bakterietyper. Benyt R-outputtet til at afgøre, om denne forskel er signifikant ved signifikansniveau $\alpha = 0.01$. (Begge dele af svaret skal være rigtige).

1 ☐

```
z1 <- replicate(10000, sample(bact1, replace = TRUE))
z2 <- replicate(10000, sample(bact2, replace = TRUE))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##        0.5%       99.5%
## -4.00000000 -0.06666667
```

Da 0 ikke er med i konfidensintervallet, er der signifikant forskel i antallet af bakteriekolonier per felt mellem de to bakterietyper.

2 ☐

```
z1 <- replicate(10000, sample(bact1, replace = TRUE))
z2 <- replicate(10000, sample(bact2, replace = TRUE))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.025, 0.975))

##       2.5%      97.5%
## -3.4666667 -0.4666667
```

Da 0 ikke er med i konfidensintervallet, er der signifikant forskel i antallet af bakteriekolonier per felt mellem de to bakterietyper.

3 ☐
```
z1 <- replicate(10000, rnorm(15, mean(bact1), sd(bact1)))
z2 <- replicate(10000, rnorm(15, mean(bact2), sd(bact2)))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##       0.5%      99.5%
## -4.0125761  0.1580288
```

Da 0 er med i konfidensintervallet, er der ikke signifikant forskel i antallet af bakteriekolonier per felt mellem de to bakterietyper.

4 ☐
```
z1 <- replicate(10000, rpois(15, mean(bact1)))
z2 <- replicate(10000, rpois(15, mean(bact2)))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##       0.5%      99.5%
## -4.1333333  0.2003333
```

Da 0 er med i konfidensintervallet, er der ikke signifikant forskel i antallet af bakteriekolonier per felt mellem de to bakterietyper.

5 ☐
```
z1 <- replicate(10000, rpois(15, mean(bact2)-mean(bact1)))
z2 <- replicate(10000, rpois(15, mean(bact2)-mean(bact1)))
d <- apply(z1, 2, mean) - apply(z2, 2, mean)
quantile(d, c(0.005, 0.995))

##      0.5%     99.5%
## -1.266667  1.266667
```

Da 0 er med i konfidensintervallet, er der ikke signifikant forskel i antallet af bakteriekolonier per felt mellem de to bakterietyper.

Fortsæt på side 27

Opgave XV

Bob kaster en dartpil mod en dartskive placeret på en væg. Lad $[X_1, X_2]$ være en bivariat normalfordelt stokastisk variabel, der måler afstanden fra centrum af skiven til hvor pilen rammer i hhv. vertikal og horisontal retning. Antag at middelværdien af både X_1 og X_2 er 0, samt at korrelationen mellem X_1 og X_2 er 0, og at $\sigma_X^2 = \sigma_Y^2 = 1$

Spørgsmål XV.1 (29)

Hvis skivens diameter er 4, hvad er så sandsynligheden for at Bob rammer væggen (Hint: start med at skrive den kvadrede afstand til centrum op)?

1 ☐ 0.02

2 ☐ 0.14

3 ☐ 0.06

4 ☐ 0.10

5 ☐ 0.18

Spørgsmål XV.2 (30)

Lise kaster ligeledes en dartpil mod dartskiven. Lises kast følger ligeledes en bivariat normal fordeling, men med variansparametrene $\sigma_X^2 = \sigma_Y^2 = \frac{1}{2}$, mens de øvrige parametre er som Bobs. Hvad er sandsynligheden for at Lises kast rammer tættere på centrum end Bobs?

1 ☐ 0.8

2 ☐ 0.75

3 ☐ 0.667

4 ☐ 0.5

5 ☐ 0.9

SÆTTET ER SLUT. God sommer!