Danmarks Tekniske Universitet *Side 1 af 23 sider.*

Skriftlig prøve: 22. juni 2017

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 7 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 for et ukorrekt svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller andet 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. 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	I.2	I.3	I.4	I.5	II.1	III.1	III.2	III.3	III.4
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

Opgave	VI.3	VI.4	VI.5	VI.6	VI.7	VI.8	VI.9	VI.10	VII.1	VII.2
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Sættet består af 23 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 de enkelte spørgsmål. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde.*

Opgave I

I et dobbeltblændet placebokontrolleret klinisk studie blev 100 personer med højt langtidsblodsukker givet en ny type medicin (betegnet 'aktiv'). Yderligere 100 personer blev givet placebo. Personerne fik medicinen over 26 uger og deres langtidsblodsukker blev bestemt ved starten af forsøget (uge 0) og efter at sidste dosis var givet (uge 26). Følgende tabel angiver gennemsnit og standard afvigelse for langtidsblodsukkeret, der måles som koncentrationen af stoffet HbA1c [%] i blodet.

Medicin	Uge	Gennemsnit	Standard afv.
Aktiv	0	8.5	1.1
Placebo	0	8.6	1.2
Aktiv	26	7.2	1.4
Placebo	26	8.1	1.5

Spørgsmål I.1 (1)

Et 95% konfidensinterval for det gennemsnitlige langtidsblodsukker efter 26 uger i gruppen, der fik den nye medicin (aktiv) er:

1 ☐ $7.2 \pm t_{0.95} \cdot \frac{1.4}{\sqrt{100}}$, hvor t-fordelingen med 99 frihedsgrader er anvendt

2 ☐ $(8.5 - 7.2) \pm t_{0.975} \cdot \frac{1.4}{10}$, hvor t-fordelingen med 99 frihedsgrader er anvendt

3 ☐ $7.2 \pm t_{0.975} \cdot \frac{1.4}{\sqrt{99}}$, hvor t-fordelingen med 99 frihedsgrader er anvendt

4 ☐ $7.2 \pm t_{0.975} \cdot 0.14$, hvor t-fordelingen med 99 frihedsgrader er anvendt

5 ☐ $(8.1 - 7.2) \pm t_{0.975} \cdot \frac{1.4}{10}$, hvor t-fordelingen med 99 frihedsgrader er anvendt

Fortsæt på side 3

Spørgsmål I.2 (2)

Når forskellen i startniveau tages med i betragtning, kan p-værdien for testet af om effekten (faldet i HbA1c) over 26 uger af den nye medicin (aktiv) adskiller sig fra placebo findes med følgende R-kode (her angiver `aktiv_week26`,`aktiv_week0`,`placebo_week26`, og `placebo_week0`, vektorer med HbA1c-niveauerne for de enkelte patienter i aktiv gruppen i uge 26 og uge 0, hhv. i placebo-gruppen i uge 26 og 0, alle sorteret efter patient ID):

1 ☐ `t.test(aktiv_week26, placebo_week26)`

2 ☐ `t.test(aktiv_week26, placebo_week26, paired=TRUE)`

3 ☐
```
t.test(aktiv_week26 - mean(c(placebo_week0, aktiv_week0)),
placebo_week26 - mean(c(placebo_week0, aktiv_week0)))
```

4 ☐ `t.test(aktiv_week26 - aktiv_week0, placebo_week26 - placebo_week0, paired=TRUE)`

5 ☐ `t.test(aktiv_week26 - aktiv_week0, placebo_week26 - placebo_week0)`

Spørgsmål I.3 (3)

Med henblik på at beregne hvor mange personer der er brug for i et nyt studie, hvor der indgår en ny gruppe personer med et middelhøjt langtidsblodsukker er der behov for at bestemme et konfidensinterval for standardafvigelsen af blodsukkersænkning blandt personer som fik den nye medicin. Antag at standardafvigelsen for blodsukkersænkning i stikprøven bestående af personer som fik den nye medicin er beregnet til 0.9.

Hvad er 95% konfidensintervallet for standardafvigelsen under disse forudsætninger?

1 ☐ $\left[\sqrt{\frac{(100-1)0.81}{128.42}},\ \sqrt{\frac{(100-1)0.81}{73.36}}\right]$

2 ☐ $\left[\sqrt{\frac{(200-1)0.81}{128.42}},\ \sqrt{\frac{(200-1)0.81}{73.36}}\right]$

3 ☐ $\left[\frac{(100-1)0.9^2}{128.42},\ \frac{(100-1)0.9^2}{73.36}\right]$

4 ☐ $\left[\sqrt{\frac{(100-1)0.81}{123.23}},\ \sqrt{\frac{(100-1)0.81}{77.05}}\right]$

5 ☐ $\left[\sqrt{\frac{(200-1)0.9^2}{239.96}},\ \sqrt{\frac{(200-1)0.9^2}{161.83}}\right]$

Fortsæt på side 4

Spørgsmål I.4 (4)

Baseret på analyserne ovenfor ønskes den nyudviklede medicin nu sammenlignet med en konkurrents medicin i et såkaldt aktiv-kontrol studie også over 26 uger. Det forventes at personer, der bliver randomiseret til gruppen, der modtager konkurrentens medicin i gennemsnit vil opleve en sænkning af blodsukkeret med 0.9 enheder (HbA1c[%]) over 26 uger. Hvorimod personer der bliver randomiseret til gruppen, der modtager den nyudviklede medicin vil opleve den samme sænkning af blodsukkeret som i det tidligere placebo-kontrollerede studie. Man antager desuden at standardafvigelsen er 1.2.

Der ønskes 95% styrke til at vise en signifikant forskel, under de specificerede antagelser, på den nyudviklede medicin og konkurrentens med et 5% signifikansniveau. Baseret på nedenstående R-kode, hvor mange personer skal der så i alt rekruteres til studiet?

```
power.t.test(delta=-0.4, sd=1.2, power = 0.95, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 234.8696
##           delta = 0.4
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
power.t.test(delta=0.9, sd=1.2, power = 0.95, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##               n = 47.18603
##           delta = 0.9
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Fortsæt på side 5

```r
power.t.test(delta=-0.4, sd=1.2, power = 0.95, sig.level = 0.05, type="paired")
```

```
##
##      Paired t test power calculation
##
##               n = 118.8917
##           delta = 0.4
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

1 ☐ 119

2 ☐ 470

3 ☐ 96

4 ☐ 48

5 ☐ 235

Fortsæt på side 6

Spørgsmål I.5 (5)

Efter udførelsen af aktiv-kontrol studiet findes nu at et 95% konfidensinterval for den estimerede behandlingseffekt målt som sænkningen i langtidsblodsukkeret (HbA1c[%]) efter 26 ugers behandling for den nyudviklede medicin i forhold til konkurrentens medicin (det vil sige for forskellen mellem behandlingsgrupperne) til $[0.33; 0.82]$.

Bedøm hvilket af følgende udsagn der er en fornuftig konklusion på studiet:

1 ☐ Den nyudviklede medicin giver en signifikant sænkning af blodsukkeret efter 26 uger, der med 95% sikkerhed er mellem 0.33% og 0.82% HbA1c.

2 ☐ Der er signifikant forskel på de to behandlinger (p-værdi < 0.05) og for en potentiel patient er der højest 5% risiko for at den nyudviklede medicin ikke sænker blodsukkeret mere end konkurrentens medicin.

3 ☐ Enhver patient kan med 95% sikkerhed forvente at den nyudviklede medicin sænker deres langtidsblodsukker (HbA1c[%]) med 0.33% - 0.82% mere end hvis de havde taget konkurrentens medicin.

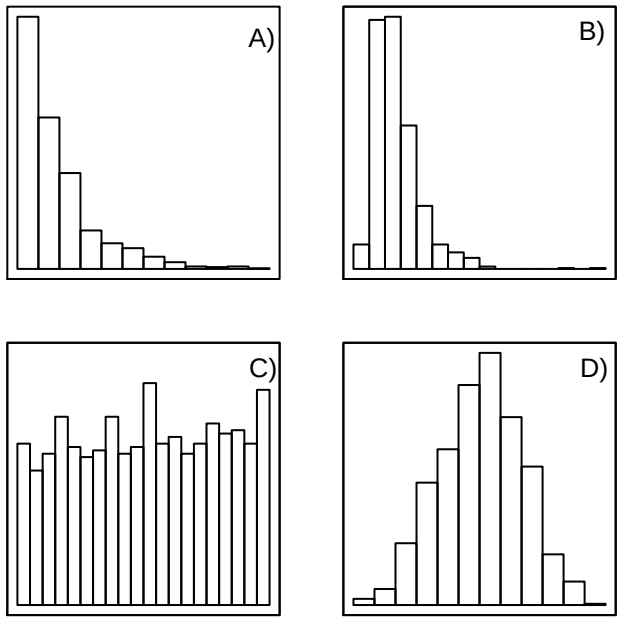
4 ☐ Der er signifikant forskel på de to behandlinger (p-værdi < 0.05) og det kan med 95% sikkerhed forventes at den nyudviklede medicin i gennemsnit sænker blodsukkeret (HbA1c[%]) med 0.33% - 0.82% mere end konkurrentens medicin.

5 ☐ Der er 95% sandsynlighed for at den nyudviklede medicin sænker blodsukkeret mere end konkurrentens medicin.

Fortsæt på side 7

Opgave II

Figuren herunder viser histogrammer af data for realisationer fra forskellige fordelinger.



Spørgsmål II.1 (6)

Hvilket af følgende udsagn omkring oprindelsen af data er antageligvis korrekt?

1 ☐ B) eksponentialfordelte data, C) normalfordelte data

2 ☐ C) exponentialfordelte data, D) normalfordelte data

3 ☐ A) exponentialfordelte data, D) uniformtfordelte data

4 ☐ A) log-normalfordelte data, B) normalfordelte data

5 ☐ A) exponentialfordelte data, B) log-normalfordelte data

Fortsæt på side 8

Opgave III

Bitterheden af vin afhænger af forskellige faktorer i vinfremstillingen. Under vinfremstillingen moses vindruerne til en blanding af juice og vindrueskaller og man kan bl.a. variere temperaturen af blandingen og om man over en periode lader vindrueskallerne være i kontakt med juicen.

I dette eksperiment har 9 bedømmere vurderet bitterheden (på en kontinuert skala) af hvidvine fremstillet under 2 forhold hvor man har varieret kontakten mellem skaller og juice (`yes`, `no`). Hver bedømmer har vurderet 4 vine under hver af de 2 forhold. Det antages at vi kan se bort fra variationen mellem bedømmere, samt at alle observationer kan antages uafhængige.

Den følgende analyse har til hensigt at undersøge hvilken effekt, hvis nogen, kontakt har på bitterheden af hvidvinene. Følgende variansanalyse er blevet udført hvor nogle af tallene er blevet maskeret:

```
> anova(lm(response ~ contact, data=wine))
Analysis of Variance Table

Response: response
          Df  Sum Sq Mean Sq F value    Pr(>F)
contact    X  3226.7       X    X     0.003479
Residuals 70  X        352.7
```

Spørgsmål III.1 (7)

Den totale kvadratafvigelsessum (SST) er:

1 ☐ 24689

2 ☐ 27563

3 ☐ 3579.4

4 ☐ 27915.7

5 ☐ 2874

Fortsæt på side 9

Spørgsmål III.2 (8)

Teststørrelsen og den kritiske værdi ($\alpha = 0.05$) for testet af om der er forskel på kontaktniveauerne er:

1 ☐ Teststørrelsen er 4.57 og den kritiske værdi er $F_{0.95}(2, 70) = 3.13$

2 ☐ Teststørrelsen er 9.149 og den kritiske værdi er $F_{0.95}(1, 70) = 3.98$

3 ☐ Teststørrelsen er 0.131 og den kritiske værdi er $F_{0.95}(1, 70) = 3.98$

4 ☐ Teststørrelsen er 9.149 og den kritiske værdi er $F_{0.975}(1, 70) = 5.25$

5 ☐ Teststørrelsen er 6.38 og den kritiske værdi er $F_{0.95}(2, 70) = 3.13$

Spørgsmål III.3 (9)

Idet gennemsnittene for de to niveauer af kontakt beregnes til $\mu_{contact:no} = 40.52$ og $\mu_{contact:yes} = 53.91$ er et 95% konfidensinterval (post-hoc) for forskellen på de 2 niveauer af kontakt:

1 ☐ $13.39 \pm 1.99\sqrt{24689(\frac{1}{36} + \frac{1}{36})}$

2 ☐ $13.39 \pm 1.96\sqrt{24689(\frac{1}{36} + \frac{1}{36})}$

3 ☐ $13.39 \pm 2.29\sqrt{352.7(\frac{1}{36} + \frac{1}{36})}$

4 ☐ $13.39 \pm 2.29\sqrt{\frac{27915.4}{70}(\frac{1}{36} + \frac{1}{36})}$

5 ☐ $13.39 \pm 1.99\sqrt{352.7(\frac{1}{36} + \frac{1}{36})}$

Spørgsmål III.4 (10)

På signifikansniveau $\alpha = 0.05$ hvad er konklusionen på studiet (både konklusion og argument skal være korrekt)?

1 ☐ Kontakt har en signifikant indflydelse på bitterheden, da $0.0035 < 0.05$.

2 ☐ Kontakt har en signifikant indflydelse på bitterheden, da $3226.7 < 70 \cdot 352.7$.

3 ☐ Kontakt har en signifikant indflydelse på bitterheden, da $3226.7 > 352.7$.

4 ☐ Kontakt har ikke en signifikant indflydelse på bitterheden, da $0.0035 < 0.05$.

5 ☐ Kontakt har ikke en signifikant indflydelse på bitterheden, da $3226.7 > 352.7$.

Fortsæt på side 10

Spørgsmål III.5 (11)

Det oplyses nu at "gentagelser" ved forskellige kontaktniveauer dækker over forskellige temperatur niveauer. Hvilke af følgende udsagn om en ny variansanalyse tabel, der tager højde for de forskellige temperaturniveauer er korrekt (SSQ er en forkortelse for "sum of squares" af Q, hvor Q kan være residualer eller behandlings effect)?

1 ☐ MS for kontakt vil forblive uændret, men p-værdien for kontakt vil ændre sig.

2 ☐ SSQ og MS og derfor også p-værdien for kontakt vil forblive uændret.

3 ☐ SSQ og MS for residualleddet vil stige men p-værdien for kontakt vil forblive uændret.

4 ☐ SSQ og MS for residualleddet vil forblive uændret.

5 ☐ SSQ og MS for både kontakt og residualleddet vil forblive uændret og derfor vil også p-værdien for kontakt forblive uændret.

Fortsæt på side 11

Opgave IV

En genstand skydes afsted (fra kordinaten $(x, y) = (0, 0)$), det antages at dens bane følger kurven for en kasteparabel uden luftmodstand. Kastelængden er givet ved udtrykket

$$x_{max} = \frac{v_0^2 \sin(2\theta)}{g}$$

hvor g er tyngdeaccelationen, θ er affyringsvinklen og v_0 er starthastigheden. Antag nu at starthastigheden og affyringsvinklen begge er usikre, mere specifikt at middelværdierne er μ_v og μ_θ for hhv. v_0 og θ, samt at varianserne er σ_v^2 og σ_θ^2. Det antages desuden at affyringsvinkel og starthastighed er uafhængige, samt at g er kendt uden usikkerhed. Til hjælp for opgaven oplyses det at $\frac{\partial \sin(z)}{\partial z} = \cos(z)$.

Spørgsmål IV.1 (12)

Hvilket af følgende udtryk er fejlophobningslovens approksimation til variansen af kastelængden?

1 ☐ $\left(\frac{2\mu_v \sin(2\mu_\theta)}{g}\right)^2 \sigma_v^2 + \left(\frac{2\mu_v^2 \cos(2\mu_\theta)}{g}\right)^2 \sigma_\theta^2$

2 ☐ $\left(\frac{2\sigma_v^2 \sin(2\sigma_\theta^2)}{g}\right)^2 \mu_v^2 + \left(\frac{2\sigma_v^2 \cos(2\sigma_\theta^2)}{g}\right)^2 \mu_\theta^2$

3 ☐ $\frac{4\mu_v \cos(2\mu_\theta)}{g}(\sigma_v^2 + \sigma_\theta^2)$

4 ☐ $\left(\frac{4\mu_v \cos(2\mu_\theta)}{g}\right)^2 (\sigma_v^2 + \sigma_\theta^2)$

5 ☐ $\frac{2\sigma_v^2 \sin(2\sigma_\theta^2)}{g}$

For specifikke fordelingsantagelser kan man konstruere simuleringsbaserede fordelinger og dermed intervaller, man antager nu at vinklen (målt i radianer) følger en uniform fordeling i intervallet $[\frac{\pi}{8}; \frac{3\pi}{8}]$ og at starthastigheden følger en normalfordeling med middelværdi 20 og standardafvigelse $\sigma_v = 2$, det antages desuden at $g = 9.81$. Til hjælp for at finde intervallet har man kørt følgende R-kode (hvoraf alt ikke nødvendigvis giver mening):

Fortsæt på side 12

```r
g <- 9.81
v0 <- 20
sigma.v <- 2
k <- 100000
theta <- runif(k, pi/8, pi * 3/8)
v <- rnorm(k, mean=v0, sd=sigma.v)
xm <- v^2 * sin(2 * theta) / g
quantile(xm, prob = c(0.005,0.01,0.025,0.975,0.99,0.995))

##     0.5%       1%     2.5%    97.5%      99%    99.5%
## 19.15289 20.43880 22.53632 54.52955 58.23360 60.63147

mean(xm) + c(-1, 1) * qnorm(0.995) * sd(xm)

## [1] 15.86690 58.26297

mean(xm) + c(-1, 1) * qnorm(0.99)  * sd(xm)

## [1] 17.92003 56.20984

mean(xm) + c(-1, 1) * qnorm(0.975) * sd(xm)

## [1] 20.93523 53.19465
```

Spørgsmål IV.2 (13)

Hvilket af følgende intervaller er det simuleringsbaserede interval (x_1, x_2) således at $P(X_{max} < x_1) = P(X_{max} > x_2)$ og $P(x_1 < X_{max} < x_2) = 0.99$?

1 ☐ $(19.15; 60.63)$

2 ☐ $(20.44; 58.23)$

3 ☐ $(17.92; 56.21)$

4 ☐ $(15.87; 58.26)$

5 ☐ $(20.94; 53.19)$

Fortsæt på side 13

Opgave V

Denne opgave handler om at lave en model til forudsigelse af CO2 udledningen fra elproduktion. I kraftværker bliver materiale afbrændt for at generere elektricitet og i forbrændingen udledes CO2. I denne forbindelse ønsker man at forudse CO2-niveauet i elproduktionen, således at man kan flytte el-forbrug til perioder med lavest CO2-niveau og dermed minimere udledningerne.

Hver dag beregnes CO2-udledningen per producerede kWh elektricitet i Danmark på baggrund af data fra ENTSO-E (European Network of Transmission system operators for electricity).

Datasættet til bestemmelse af en god model til forudsigelse af CO2 består af gennemsnitlige timeværdier af følgende variabler, hvoraf de sidste tre er 24 timers prognoser:

Variabel	Beskrivelse	Range	Enhed
`co2intensity`	CO2 intensiteten	$[113, 566]$	gCO2eq/kW
`windspeed`	Vindhastighed	$[1.7, 11.4]$	m/s
`importDE`	El-import fra Tyskland	$[-2300, 1845]$	kW
`generation`	El-produktion	$[920, 3910]$	kW

Data er indlæst i R i data.table `X`.

Der er fittes en lineær regressionsmodel med et intercept og de tre forklarende variabler

$$Y_{\text{co2intensity},i} = \beta_0 + \beta_{\text{w}} X_{\text{windspeed},i} + \beta_{\text{im}} X_{\text{importDE},i} + \beta_{\text{g}} X_{\text{generation},i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

og resultatet skrives ud:

```
##
## Call:
## lm(formula = "co2intensity ~ windspeed + importDE + generation",
##     data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.232 -26.026  -5.613  19.376 153.485
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 263.146473   8.542834   30.80   <2e-16 ***
## windspeed   -17.202953   0.761038  -22.61   <2e-16 ***
## importDE     -0.031120   0.001192  -26.11   <2e-16 ***
## generation    0.076791   0.002926   26.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.73 on 721 degrees of freedom
## Multiple R-squared:  0.7855,Adjusted R-squared:  0.7846
## F-statistic:   880 on 3 and 721 DF,  p-value: < 2.2e-16
```

Vi antager i det følgende at uafhængighedsantagelsen er opfyldt.

Spørgsmål V.1 (14)

Ud fra den sædvanligt anvendte procedure til udvælgelse af model (og signifikansniveau $\alpha = 0.05$), bør denne model så reduceres (både konklusion og argument skal være korrekt)?

1 ☐ Ja, da den forklarede varians er højere end signifikansniveauet

2 ☐ Nej, da den forklarede varians er højere end det valgte signifikansniveau

3 ☐ Ja, da $\hat{\sigma}/df = 39.73/721 = 0.0551$ er højere end det valgte signifikansniveau

4 ☐ Nej, da $\hat{\sigma}/df = 39.73/721 = 0.0551$ er højere end det valgte signifikansniveau

5 ☐ Nej, da samtlige kofficienter i modellen er signifikant forskellige fra nul med det valgte signifikansniveau

Spørgsmål V.2 (15)

Hvilket af følgende udsagn er ikke korrekt (det usande udsagn skal udpeges)?

1 ☐ Modellens prædiktion af CO2-intensitet ved vindstille, med ingen el-import fra Tyskland og en el-produktion på 2000 kW, er 417 gCO2eq/kW

2 ☐ Standard afvigelsen af afvigelserne er estimeret til 39.73

3 ☐ Modellen har forklaret 79% af variansen

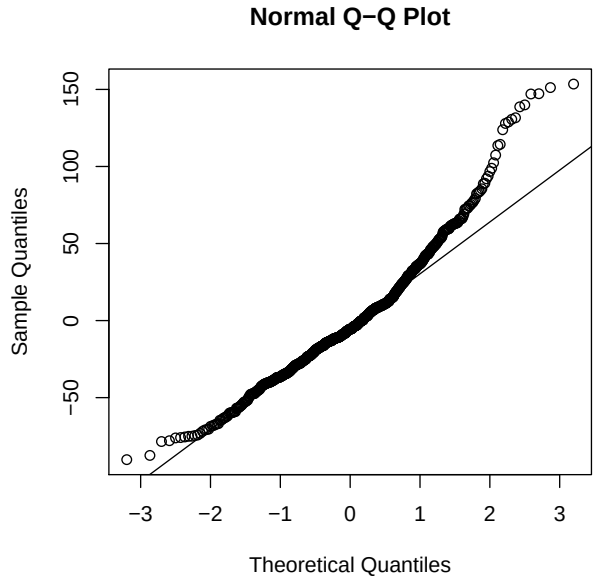
4 ☐ Den vigtigste af de tre forklarende variabler er vindhastigheden da $|\hat{\beta}_{\text{im}}| < |\hat{\beta}_{\text{g}}| < |\hat{\beta}_{\text{w}}|$

5 ☐ Det er estimeret at når vindhastigheden stiger, så falder CO2-intensiteten

Fortsæt på side 15

Spørgsmål V.3 (16)

I forbindelse med validering af modellen, laves bl.a. et qq-normalplot af residualerne:



Hvilken af følgende vurderinger kan drages på baggrund af dette plot?

1 ☐ Antagelsen om at afvigelserne ε_i er i.i.d. er ikke opfyldt

2 ☐ Antagelsen om at afvigelserne ε_i er i.i.d. er opfyldt

3 ☐ Antagelsen om at afvigelserne ε_i er normalfordelt er ikke opfyldt

4 ☐ Antagelsen om at afvigelserne ε_i er normalfordelt er opfyldt

5 ☐ Det kan afvises at der er en lineær relation mellem CO2-intensiteten og de forklarende variabler

Fortsæt på side 16

Spørgsmål V.4 (17)

Ved matrixformuleringen af en lineær regressionsmodel

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

indgår den såkaldte designmatrix $\boldsymbol{X}$. I den anvendte model og det anvendte datasæt hvad bliver da dimensionerne af denne?

1 ☐ 2884 rækker og 4 kolonner

2 ☐ 725 rækker og 4 kolonner

3 ☐ 4 rækker og 4 kolonner

4 ☐ 5 rækker og 4 kolonner

5 ☐ 4 rækker og 5 kolonner

Spørgsmål V.5 (18)

Nu vil man beregne en prædiktion af et nyt punkt. Hvis de nye prognoseværdier er i vektoren $\boldsymbol{x}_{\text{new}}$, og modellen er skrevet på matrixform, som under forrige spørgsmål. Hvorledes beregnes prædiktionen af CO2-intensiteten da?

1 ☐ $\hat{Y}_{\text{co2intensity,new}} = \boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

2 ☐ $\hat{Y}_{\text{co2intensity,new}} = \boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}} + \varepsilon_{\text{new}}$

3 ☐ $\hat{Y}_{\text{co2intensity,new}} = \text{V}(\boldsymbol{x}_{\text{new}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} + \varepsilon_{\text{new}})$

4 ☐ $\hat{Y}_{\text{co2intensity,new}} = \text{V}(\boldsymbol{x}_{\text{new}}\hat{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}\boldsymbol{X}^T\boldsymbol{Y})$

5 ☐ $\hat{Y}_{\text{co2intensity,new}} = \text{V}(\boldsymbol{x}_{\text{new}}\hat{\boldsymbol{\beta}})$

Fortsæt på side 17

Opgave VI

I en produktionslinie hvor der produceres låg ønsker man at undersøge hvad sandsynligheden for fejl er. Til dette formål har man udtaget 229 prøver med 770 låg i hver prøve. Tabellen herunder viser resultatet af prøvetagningen.

Antal defekte	0	1	2	3	4	5	6	7	8	9
Antal prøver	131	38	28	11	4	5	5	2	3	2

Således er der eksempelvis observeret 38 prøver (med hver 770 låg) med et defekt låg i hver. Hvis man antager at sandsynligheden for fejl er den samme i hver af de 229 prøver kan estimatet af sandsynligheden for fejl således udregnes til

$$\begin{aligned}\hat{p} =& \frac{38 + 2 \cdot 28 + 3 \cdot 11 + 4 \cdot 4 + 5 \cdot 5 + 6 \cdot 5 + 7 \cdot 2 + 8 \cdot 3 + 9 \cdot 2}{770 \cdot 229} \\ =& 0.00144\end{aligned}$$

Spørgsmål VI.1 (19)

Hvad er det gennemsnitlige antal defekte låg pr. prøve?

1 ☐ 4.5

2 ☐ 22.9

3 ☐ 0.33

4 ☐ 1.109

5 ☐ 254

Spørgsmål VI.2 (20)

Hvis `defective` og `samples` betegner hhv. første og anden række i tabellen herover og `m` betegner det empiriske gennemsnit af antallet af defekte låg per prøve, hvilken af følgende R-kommandoer udregner da den empiriske varians for antallet af defekte låg per prøve?

1 ☐ `sum((defective - m)^ 2 * samples)/ (229 * 770)`

2 ☐ `sum((defective - m)^ 2) / 9`

3 ☐ `sqrt(sum((defective - m)^ 2 * samples)/ (229 * 770))`

4 ☐ `sum((defective - m)^ 2 * samples) / 228`

5 ☐ `sum((samples - m)^ 2) / 9`

Fortsæt på side 18

Spørgsmål VI.3 (21)

Hvilket interval bliver IQR (inter quartile range) for antallet af defekte låg per prøve regnet ud fra?

1 ☐ $[3; 28]$

2 ☐ $[2.25; 6.75]$

3 ☐ $[0; 2]$

4 ☐ $[3.25; 23.75]$

5 ☐ $[2; 7]$

Spørgsmål VI.4 (22)

Hvis man antager at sandsynligheden for at et låg er defekt er $p = 0.00144$ hvad er da forventede antal prøver med 5 defekte låg når der i alt er 229 prøver med hver 770 låg?

1 ☐ 0.00458

2 ☐ 0.0000225

3 ☐ 0.0173

4 ☐ 1.05

5 ☐ 3.53

Spørgsmål VI.5 (23)

Hvad er et 95% konfidensinterval for sandsynligheden for fejl på et enkelt låg?

1 ☐ $0.00144 \pm z_{0.975}\sqrt{\frac{0.00144(1-0.00144)}{770 \cdot 229}} = [0.00126; 0.00162]$

2 ☐ $0.00144 \cdot 229 \pm z_{0.975}\sqrt{\frac{0.00144(1-0.00144)\cdot 229}{770}} = [0.289; 0.370]$

3 ☐ $\frac{229}{770} \pm z_{0.975}\sqrt{\frac{229/770(1-229/770)}{770}} = [0.265; 0.330]$

4 ☐ $0.00144 \cdot \frac{770}{229} \pm z_{0.975}\sqrt{\frac{0.00144(1-0.00144)}{770 \cdot 229}} = [0.00466; 0.00502]$

5 ☐ $0.00144 \pm z_{0.975}\frac{\sqrt{229/770(1-229/770)}}{770} = [0.000276; 0.00260]$

Fortsæt på side 19

Spørgsmål VI.6 (24)

Man planlægger nu en ny prøvetagning. Hvor mange prøver med hver 770 låg skal der udtages hvis man ønsker en maksimal fejl på 0.0001 med signifikansniveau $\alpha = 0.05$ og man benytter den observerede andel af fejl (0.00144) som scenarie?

1 ☐ 384

2 ☐ 2413

3 ☐ 545

4 ☐ 506

5 ☐ 718

Man beslutter nu at fordelingsantagelsen er tvivlsom og ønsker derfor at konstruere et simulationsbaseret 95% konfidensinterval for defektsandsynligheden, til formålet har man kørt følgende R-kode (hvoraf alt ikke nødvendigvis giver mening), i koden angiver `new.samp` en vektor med 131 nuller, 38 et-taller osv.

```
k <- 100000
defective <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
samples <- c(131, 38, 28, 11, 4, 5, 5, 2, 3, 2)
new.samp <- rep(defective, samples)

sim <- replicate(k, sample(new.samp, replace = TRUE))

sim.p <- apply(sim, 2, sum) / (229 * 770)

quantile(sim.p, c(0.025, 0.05, 0.95, 0.975))

##        2.5%          5%         95%       97.5%
## 0.001139908 0.001185278 0.001707027 0.001758067

c(mean(sim.p), sd(sim.p))

## [1] 0.0014406186 0.0001587641
```

Fortsæt på side 20

```
################################################
##
################################################
sim2 <- replicate(k, sample(samples, replace=TRUE))

sim.p2 <- apply(sim2, 2, sum) / 770

quantile(sim.p2, c(0.025, 0.05, 0.95, 0.975))

##       2.5%         5%        95%      97.5%
## 0.07012987 0.08831169 0.58961039 0.63896104

c(mean(sim.p2), sd(sim.p2))

## [1] 0.2965242 0.1553125
```

Spørgsmål VI.7 (25)

Det simulationsbaserede konfideninterval bliver:

1 ☐ $[0.00114; 0.00176]$

2 ☐ $[0.0883; 0.590]$

3 ☐ $[0.041; 0.552]$

4 ☐ $[0.0701; 0.639]$

5 ☐ $[0.00119; 0.00171]$

Man udtager nu en tilsvarende prøve for en anden produktionslinje (dvs. 229 prøver med 770 låg i hver), resultatet ses i tabellen herunder

Antal defekte	0	1	2	3	4	5	6	7	8	9
Antal prøver	73	84	46	17	5	3	1	0	0	0

Estimatet for defektsandsynligehen kan udregnes på tilsvarende vis som ovenfor og resultatet bliver $\hat{p}_2 = 0.00152$. Man ønsker nu at undersøge om der er forskel på defektsandsynligheden i 2 produktionslinjer.

Fortsæt på side 21

Spørgsmål VI.8 (26)

Hvad bliver den sædvanlig teststørrelse og konklusion for forskellen i defektsandsynligheden i de 2 produktionslinier (idet der benyttes signifikansniveau $\alpha = 0.05$)?

1 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{229\cdot 770}}} = 0.618$, og der kan ikke påvises en forskel da $|0.618| < 1.96$

2 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{229}}} = 0.0223$, og der er signifikant forskel da $0.0223 < 0.05$

3 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{770}}} = 0.0408$, og der er signifikant forskel da $0.0408 < 0.05$

4 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\left(\frac{1}{229}+\frac{1}{770}\right)}} = 0.0276$, og der kan ikke påvises en forskel da $|0.0276| < 1.96$

5 ☐ $\frac{0.00152-0.00144}{\sqrt{0.00148(1-0.00148)\cdot\frac{2}{229}}} = 0.0223$, og der kan ikke påvises en forskel da $|0.0223| < 1.96$

Uanset udfaldet af forrige spørgsmål beslutter man at lave en mere generel sammenligning af de 2 fordelinger ved hjælp af test i antalstabeller. Til formålet har man opstillet nedenstående tabel

Defective	line 1	line 2
0	131	73
1	38	84
2	28	46
3	11	17
$4-5$	9	8
>5	12	1

Fortsæt på side 22

Spørgsmål VI.9 (27)

Hvad er bidraget til den sædvanlige teststørrelse for gruppen af 4-5 defekte (summen af bidragene fra begge produktionslinier)?

1 ☐ $\frac{1}{17}$

2 ☐ $\frac{9^2}{8.5} + \frac{8^2}{8.5}$

3 ☐ $\frac{2}{17}$

4 ☐ $\frac{8}{8.5} + \frac{9}{8.5}$

5 ☐ $\frac{1}{34}$

Spørgsmål VI.10 (28)

Den sædvanlige teststørrelse er nu udregnet til 48.865, hvad bliver p-værdi og konklusion for testen om de to fordelinger er ens (brug konfidensniveau $\alpha = 0.05$)?

1 ☐ P-værdien bliver $4.31 \cdot 10^{-7}$, og det kan ikke afvises at fordelingerne er ens

2 ☐ P-værdien bliver $2.36 \cdot 10^{-9}$, og det afvises at fordelingerne er ens

3 ☐ P-værdien bliver 0.0201, og det kan ikke afvises at fordelingerne er ens

4 ☐ P-værdien bliver $4.31 \cdot 10^{-7}$, og det afvises at fordelingerne er ens

5 ☐ P-værdien bliver 0.0201, og det afvises at fordelingerne er ens

Fortsæt på side 23

Opgave VII

Lad $X \sim N(\mu_1, \sigma_1^2)$ og $Y \sim N(\mu_2, \sigma_2^2)$, lad desuden X og Y være uafhængige.

Spørgsmål VII.1 (29)

Hvis vi antager at $\mu_1 = \mu_2 = 0$, hvad er så $Var(X^2 + Y^2)$?

1 ☐ $2\sigma_1^4 + 2\sigma_2^4$

2 ☐ $4\sigma_1^2 + 4\sigma_1^2$

3 ☐ $4\sigma_1^2 + 4\sigma_1^2 + \sigma_1^2\sigma_2^2$

4 ☐ $\sigma_1^4 + \sigma_2^4$

5 ☐ $2\sigma_1^2 + 2\sigma_1^2$

Spørgsmål VII.2 (30)

Hvis vi nu antager at $\sigma_1^2 = 2\sigma_2^2$, hvad er $P((X - \mu_1)^2 > (Y - \mu_2)^2)$ så?

1 ☐ $P(Z > 0)$, hvor Z følger en χ^2-fordeling med 1 frihedsgrad.

2 ☐ $P\left(Z > \frac{1}{2}\right)$, hvor Z følger en F-fordeling med 1 og 1 frihedsgrader.

3 ☐ $P\left(Z > \frac{1}{2}\right)$, hvor Z følger en t-fordeling med 2 frihedsgrader.

4 ☐ $P(Z > 2)$, hvor Z følger en χ^2-fordeling med 1 frihedsgrad.

5 ☐ $P\left(|Z| < \frac{1}{2}\right)$, hvor Z følger en t-fordeling med 1 frihedsgrad.

SÆTTET ER SLUT. God sommer!