Universitetet i Stavanger

PRØVEEKSAMEN I: STA100 SANNSYNLIGHETSREGNING OG STATISTIKK

VARIGHET: 5 TIMER DATO: 1. APRIL 2020

EKSAMEN BESTÅR AV 4 OPPGAVER PÅ 6 SIDER.

HJELPEMIDLER: Alle tekniske hjelpemidler er lovlige. Det er *ikke* lov å få hjelp av andre personer i arbeidet med eksamensoppgaven.

MERK ANGÅENDE BESVARELSEN: Det er viktig at du skriver besvarelsen på en slik måte at den som leser besvarelsen din skjønner hva du har tenkt og gjort. Skriv ryddig og oversiktelig, og bruk notasjon og skrivemåte lært i kurset. Det er viktig at du skriver nok til å få frem hele resonnementet ditt, ikke bare sluttsvaret.

NB! Eksamensforskriften gjelder også ved hjemme-eksamen. Det er ikke lov å få hjelp av andre i arbeidet med eksamensoppgaven. På første side i besvarelsen skal du skrive følgende tekst:

Jeg bekrefter å ha fulgt bestemmelsene i eksamensforskriften. Jeg er innforstått med at fusk eller forsøk på fusk vil bli sanksjonert.

Oppgave 1

Lengden på laks som går opp i ei bestemt elv for å gyte kan antas å være normalfordelt med forventet lengde $\mu = 90$ cm og standardavvik $\sigma = 10$ cm. Lengdene på ulike lakser er uavhengige.

a) Regn ut sannsynligheten for at en laks er over 100 cm.

Finn den lengden a som er slik at kun 1% av laksene er lengre enn a cm.

Laksefiskere i denne elva har en sesongkvote på 5 laks. Hanna er ivrig laksefisker og fisker til hun har fylt sesongkvoten. Regn ut sannsynligheten for at gjennomsnittslengden for de 5 laksene er over 100 cm.

I et større vassdrag er det to laksestammer, stamme A og stamme B. For stamme A er forventet lengde $\mu_A = 90$ og standardavviket $\sigma_A = 10$, mens for stamme B er $\mu_B = 96$ og $\sigma_B = 8$. Lengdene er som før uavhengige og normalfordelte.

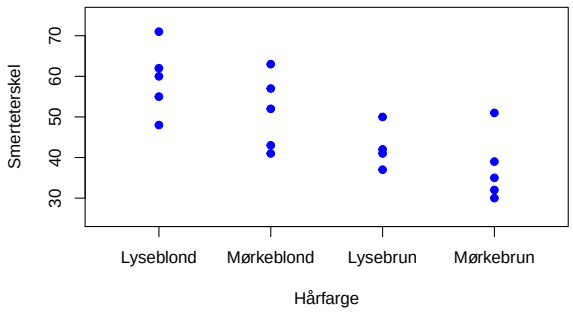
Det er videre kjent at 70% av laksen i vassdraget er fra stamme A, mens 30% er fra stamme B.

b) Regn ut sannsynligheten for at en laks fra vassdraget er over 100 cm.

Regn ut sannsynligheten for at en laks fra stamme A er lengre enn en laks fra stamme B.

Oppgave 2

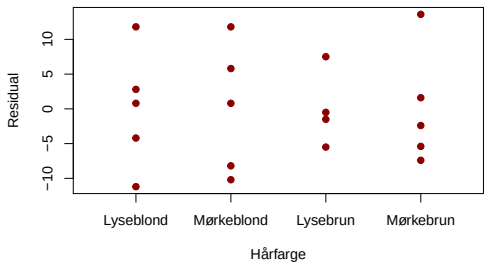
I en studie har man målt ulike personers smerteterskel og blant annet undersøkt om det er en sammenheng mellom smerteterskel og hårfarge. (Jo høyere smerteterskel jo mer smerte tåler man.) Plott av dataene (fra: `www.statsci.org/data/oz/blonds.html`), samt en tabell med gjennomsnitt ($\bar{x}$), utvalgsstandardavvik (s) og antall målinger (n) for smerteterskel for hver hårfarge er gitt under.

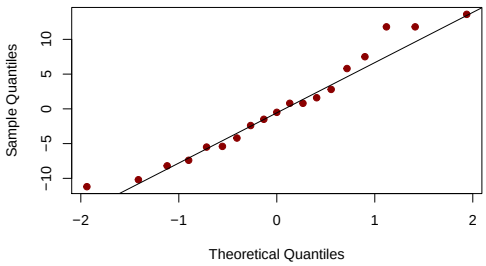


Hårfarge	$\bar{x}$	s	n
Lyseblond	59.2	8.53	5
Mørkeblond	51.2	9.28	5
Lysebrun	42.5	5.45	4
Mørkebrun	37.4	8.32	5

R-utskrifter og residualplott for en enveis variansanalyse er gitt under.

```
> vmod = aov(smerte~as.factor(farge),data=smerte)
> summary(vmod)
                  Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(farge)   3   1361   453.6   6.791 0.00411
Residuals         15   1002    66.8
---
```





a) Sett opp modellen og antagelsene for enveis variansanalyse og gi en praktisk fortolkning av hvert element i modellen i denne konkrete situasjonen.

Er de forskjell i smerteterskel mellom personer med ulike hårfarge? Formuler problemstillingen som en hypotesetest og finn utfallet av testen. Bruk 5% nivå. Forklar hva utfallet av testen betyr i praksis.

Hvilke antagelsen kan vi sjekke ut fra de to residualplottene? Ser disse antagelsene ut til å være oppfylte? Forklar hvorfor/hvorfor ikke.

b) Vi skal nå spesielt sammenligne personer med mørkeblondt og mørkebrunt hår. Kan vi konkludere med at det er forskjell i smerteterskel mellom personer med mørkeblondt og personer med mørkebrunt hår? Formuler problemstillingen som en hypotesetest og utfør testen. Bruk 5% signfikansnivå.

Spesifiser hvilke antagelser som ligger til grunn for testen du utfører.

Oppgave 3

Levetiden til en bestemt type lysrør er eksponentialfordelt med forventning 1.6 år. La T være levetiden til et tilfeldig lysrør. Anta at levetidene til ulike lysrør er uavhengige.

a) Vis at $P(T > 1) = 0.535$.

Finn $P(T < 1.6)$.

I et rom installeres 8 lysrør av den aktuelle typen. Finn sannsynligheten for at minst 6 av disse lysrørene fremdeles fungerer etter ett år.

I en bygning er det installert 72 lysrør av den aktuelle typen. Når et lysrør feiler blir det erstattet av et nytt lysrør. Det kan da vises at antall lysrør som feiler i løpet av t år er Poisson-fordelt med intensitet $\lambda = 72/1.6 = 45$ per år.

b) Finn sannsynligheten for at minst tre lysrør feilet i løpet av en måned ($t = 1/12$).

Finn sannsynligheten for at minst 36 lysrør feiler i løpet av ett år.

Sannsynlighetstettheten for eksponentialfordelingen formulert med forventningsverdien β som parameter kan skrives:

$$f(t) = (1/\beta)e^{-t/\beta}, \text{ for } t \geq 0,$$

Sammenhengen med formuleringen i boka/forelesningsnotatene er at $\beta = 1/\lambda$ og vi får dermed at $\mathrm{E}(T) = \beta$.

For en ny variant av lysrørene er forventet levetid β ukjent. For å estimere β registerer man for n uavhengige lysrør om de fremdeles fungerer etter ett år eller ikke.

La $p = P(T > 1)$.

c) Finn et estimat og et tilnærmet 95% konfidensintervall for p når man observerte at 73 av 100 lysrør fungerte etter ett år.

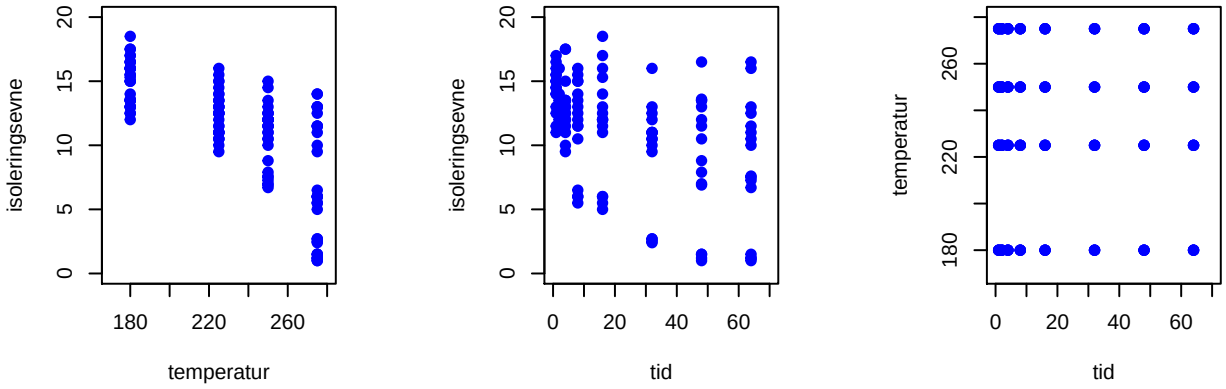
Vis at $p = e^{-1/\beta}$.

Ta utgangspunkt i konfidensintervallet for p over og finn et tilnærmet 95% konfidensintervall for β.

Hva vil være en ulempe med dette konfidensintervallet for β fremfor et intervall basert på registrering av de eksakte levetidene til alle lysrørene?

Oppgave 4

I denne oppgaven skal vi jobbe med data fra et eksperiment hvor man undersøkte reduksjon i isoleringsevne for elektriske isolatorer som ble utsatt for belastning med ulike kombinasjoner av temperatur og tid. Responsen er isoleringsevne i kilovolt, temperatur måles i grader Celcius og tid i antall uker. Det er totalt gjort $n = 128$ målinger. Plott av dataene er vist under.



Først skal vi se på en enkel lineær regresjonsmodell med isoleringsevne som responsvariabel (Y-variabel) og temperatur som forklaringsvariabel (x-variabel). En R-utskrift for denne modellen er gitt under.

```
> regmod1 = lm(isoleringsevne~temperatur, data=isolatordata)
> summary(regmod1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.511906   1.767079   17.27   <2e-16 ***
temperatur  -0.082898   0.007515  -11.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.983 on 126 degrees of freedom
Multiple R-squared:  0.4913,Adjusted R-squared:  0.4872
F-statistic: 121.7 on 1 and 126 DF,  p-value: < 2.2e-16
```

a) Hvorfor er det rimelig å bruke isoleringsevne som Y-variabel og temperatur som x-variabel, og ikke motsatt, når vi tilpasser en regresjonsmodell mellom disse to variablene?

Skriv ned den estimerte regresjonslinja, og gi en praktisk tolkning av det estimerte stigningstallet.

Forklar prinsippet som brukes for å finne den estimerte regresjonslinja. (Du trenger ikke gi utledning av formler, bare forklar prinsippet som ligger til grunn.)

Bruk informasjon i datautskriften til å finne korrelasjonen mellom temperatur og isoleringsevne.

I resten av oppgaven skal vi se på en multippel regresjonsmodell hvor både temperatur og tid tas med som forklaringsvariabler, dvs modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

der Y er isoleringsevne, x_1 er temperatur og x_2 er tid. Vi antar som vanlig at $e \sim N(0, \sigma)$ og at for ulike målinger så er $e_1, \ldots, e_n$ uavhengige.

En R-utskrift for denne modellen er gitt under.

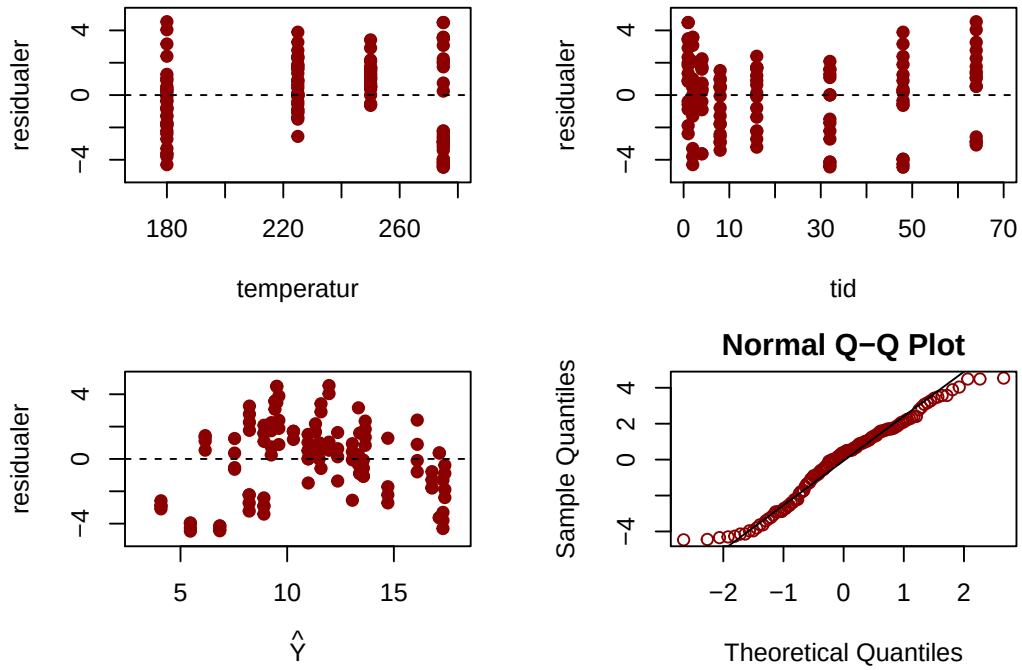
```
> regmod2 = lm(isoleringsevne~temperatur+tid, data=isolatordata)
> summary(regmod2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.396355   1.369447  23.657  < 2e-16 ***
temperatur  -0.082898   0.005762 -14.387  < 2e-16 ***
tid         -0.086146   0.009114  -9.452 2.52e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.287 on 125 degrees of freedom
Multiple R-squared:  0.7033,Adjusted R-squared:  0.6986
F-statistic: 148.2 on 2 and 125 DF,  p-value: < 2.2e-16
```

b) Sammenlign det estimerte standardavviket (estimert σ) i de to modellene (modellen med bare temperatur og modellen med både temperatur og tid) og kommenter. Hvorfor er det rimelig at verdien har endret seg i den retningen vi ser?

Sammenligne også R^2 og R^2-justert i de to modellene og kommenter. Hvilken modell ser ut for å være best fra disse kriteriene?

Det har tidligere vært vanlig å gå ut fra at isoleringsevnen reduseres med 1 for hver tiende uke (eller dermed med 0.1 per uke), men det påstås nå at reduksjonen ikke går så fort. Formuler denne problemstillingen som en hypotesetest om β_2 (der nullhypotesen er en reduksjon på 0.1 per uke). Utfør testen. Bruk 5% signifikansnivå.
(Hint: Følgende kvantiler i t-fordelingen oppgis: $t_{0.1,125} = 1.288$, $t_{0.05,125} = 1.657$ og $t_{0.025,125} = 1.979$.

Ulike plott av residualene til regresjonsmodellen er gitt under.



c) Hvordan er residualet til en observasjon i denne regresjonsmodellen definert? Regn ut residualet til en måling hvor isoleringsevne er 13.6, temperatur er 180 og tid er 48.

Forklar hvilke modelleantagelser vi kan sjekke fra hvert av de fire residualplottene. Diskuter hvorvidt disse antagelsene ser ut for å være oppfylte eller ikke. Gi forslag til mulige forbedringer av modellen.

Hvilket annet plott av residualene enn de vist kunne vært relevant å se på og hvilken modellantagelse kunne vi sjekket fra dette plottet?