EKSAMEN I: STA100 SANNSYNLIGHETSREGNING OG STATISTIKK

VARIGHET: 4 TIMER DATO: 14. MAI 2018

HJELPEMIDLER: Godkjent enkel kalkulator.

EKSAMEN BESTÅR AV 4 OPPGAVER PÅ 5 SIDER OG 5 SIDER VEDLEGG, TOTALT 10 SIDER.

EMNEANSVARLIG: Jan Terje Kvaløy TELEFON: 51 83 22 55

---

## Oppgave 1

En fabrikk produserer en type maskindel hvor det er kjent at 8% av de produserte delene har en indre feil som på sikt kan føre til at delen svikter. Det er uavhengig fra del til del om de har en slik indre feil eller ikke.

La $X$ være antall deler blant $n$ tilfeldig valgte deler som har indre feil.

a) Hvilken fordeling har $X$? Begrunn svaret.

En kunde kjøper $n = 20$ maskindeler. Hva er sannsynligheten for at minst 2 av disse har indre feil?

En annen kunde kjøper $n = 200$ maskindeler. Hva er sannsynligheten for at minst 20 av disse har indre feil?

Fabrikken mottar jevnlig klager fra kunder over deler som svikter etter en tids bruk på grunn av den indre feilen. For å prøve å oppdage og luke ut maskindeler med indre feil begynner fabrikken med å gjøre en ultralydtest av alle produserte deler. Denne ultralydtesten kan ofte avsløre indre feil, men testen er ikke helt perfekt. Noen ganger overser testen feil og andre ganger sier testen at der er en feil uten at det er tilfelle. Dersom en del har feil er sannsynligheten 96% for at testen oppdager det. Dersom en del ikke har feil er det 6% sannsynlighet for at testen likevel indikerer feil. Som før har 8% av de produserte delene feil.

La $F$ være hendelsen at en enhet har en feil, og la $U$ være hendelsen at ultralydtesten indikerer feil.

b) Ut fra opplysningene gitt over, skriv ned hva $P(F)$, $P(U|F)$ og $P(U|\overline{F})$ er.

Regn ut $P(\overline{F})$ og $P(\overline{U}|\overline{F})$.

Vis at sannsynligheten for at testen indikerer feil på en tilfeldig del er $P(U) = 0.132$.

Gitt at testen ikke indikerer feil på en del, hva er sannsynligheten for at delen likevel har feil?

Fabrikken er interessert i å følge med på om andelen produserte deler hvor ultralydtesten indikerer feil er stabil over tid. De bestemmer seg derfor for å lage et kontrolldiagram for $p = P(U)$, der det er kjent at i en stabil situasjon så er $p = 0.132$ som regnet ut i punkt b).

For hver gang 100 nye deler er blitt testet beregnes andelen blant de 100 delene hvor ultralydtesten indikerer feil, og denne andelen plottes inn i kontrolldiagrammet.

c) Regn ut kontrollgrensene for et $\hat{p}$-diagram for andel deler hvor ultralydtesten indikerer feil.

Resultatet av de tolv første utregningene av andel deler hvor testen indikerer feil i løpet av 100 tester ble:

0.07 0.13 0.25 0.11 0.10 0.17 0.12 0.13 0.15 0.08 0.18 0.17

Tegn $\hat{p}$-diagrammet for dataene og kommenter resultatet.

Hva vil den praktiske tolkningen av en alarm over øvre kontrollgrense være i denne situasjonen?

## Oppgave 2

Forekomsten av jordskjelv i Norge av en slik styrke at de kan merkes av mennesker kan beskrives som en Poisson-prosess med $\lambda = 10$ per år. Dvs antall jordskjelv per år er Poisson-fordelt med forventningsverdi 10.

a) Finn sannsynligheten for 8 jordskjelv i løpet av ett år.

Finn sannsynligheten for mer enn 4 jordskjelv i løpet av et halvt år.

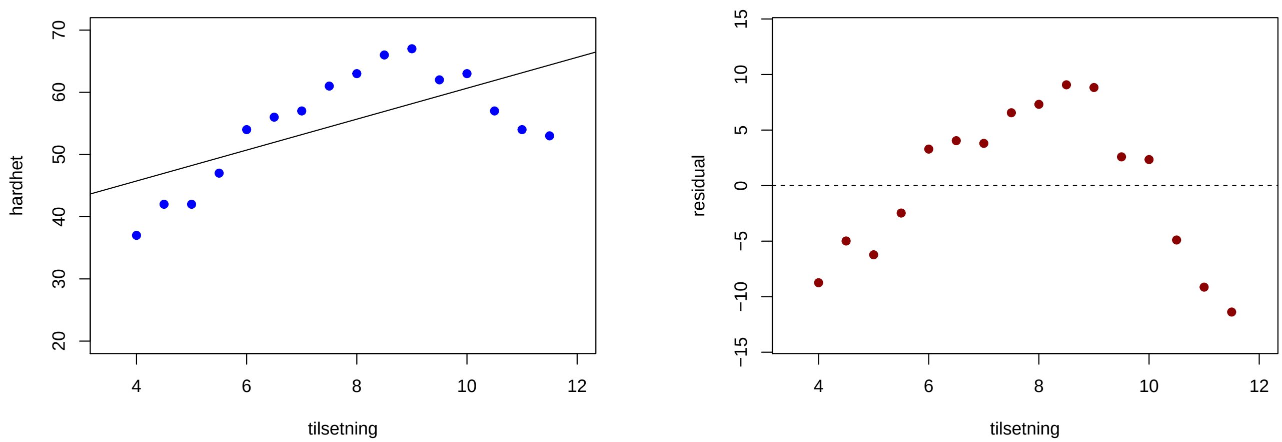
Finn sannsynligheten for at det går mer enn to måneder mellom to etterfølgende jordskjelv.

Oppgave 3

I fremstilling av en ny legering er det eksperimentert med mengden av et tilsetningsstoff som har betydning for legeringen sin hardhet. Data over mengde tilsetningsstoff og hardhet til legeringen for 16 legeringsprøver er gitt i tabellen under.

| tilsetning, $x_i$ | 4 | 4.5 | 5 | 5.5 | 6 | 6.5 | 7 | 7.5 | 8 | 8.5 | 9 | 9.5 | 10 | 10.5 | 11 | 11.5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hardhet, $y_i$ | 37 | 42 | 42 | 47 | 54 | 56 | 57 | 61 | 63 | 66 | 67 | 62 | 63 | 57 | 54 | 53 |

To plott og deler av en datautskrift fra R for en enkel lineær regresjonsmodell tilpasset til dataene er vist under.




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.8015     6.1743   5.798 4.62e-05 ***
tilsetning    2.4853     0.7636   3.255  0.00576 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.04 on 14 degrees of freedom
Multiple R-squared:  0.4307,Adjusted R-squared:  0.3901
F-statistic: 10.59 on 1 and 14 DF,  p-value: 0.00576
```

a) Skriv ned formelen og de vanlige antagelsene for den generelle enkle lineære regresjonsmodellen.

Hvorfor er det rimelig å bruke tilsetning som $x$-variabelen og hardhet som $Y$-variabelen når vi tilpasser en regresjonsmodell til disse dataene?

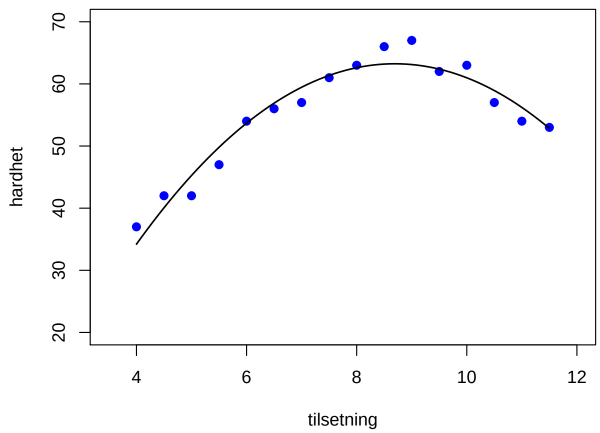
Hvordan er residualet til en observasjon definert? Regn ut residualet til den første observasjonen ($x_1 = 4,\ y_1 = 37$).

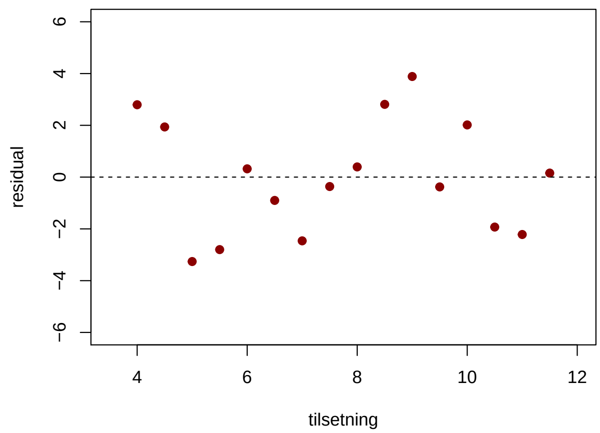
Hvilke antagelser om modellen kan vi sjekke ut fra plottet av residualer gitt over? Ser disse antagelsene ut til å være oppfylte?

I stedet for den enkle lineære modellen skal vi nå prøve en modell med en andregradskurve:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

der $e \sim N(0, \sigma)$ og vi antar at $e_1, \ldots, e_n$ er uavhengige. Plott og datautskrift er vist under.




```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -36.3934     7.1828  -5.067 0.000216 ***
tilsetning       22.9240     1.9632  11.677 2.90e-08 ***
I(tilsetning^2)  -1.3186     0.1256 -10.501 1.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.373 on 13 degrees of freedom
Multiple R-squared:   0.94,Adjusted R-squared:  0.9307
F-statistic: 101.8 on 2 and 13 DF,  p-value: 1.147e-08
```

b) Skriv ned den estimerte regresjonskurven.

Regn ut residualet til den første observasjonen ($x_1 = 4,\ y_1 = 37$) i denne modellen.

Tyder plottene over på at modellen med andregradsledd er bedre enn den lineære modellen på forrige side? Forklar kort.

Bruk den estimerte regresjonskurven til å anslå hvilken mengde tilsetningsstoff som gir høyest hardhet.

Trengs andregradsleddet i modellen? Formuler dette som en hypotesetest og gi resultatet av testen.

Oppgave 4

Sammen med oljen som pumpes opp fra oljebrønner kommer en del vann som skilles fra oljen. Dette vannet kalles produksjonsvann, og det tas jevnlig en rekker prøver av dette vannet noe som både gir viktig informasjon om reservoaret og er viktig for miljøovervåkning. I denne oppgaven skal vi se på konsentrasjonen av grunnstoffet brom i produksjonsvannet.

Vi skal i hele denne oppgaven anta at med de målemetodene som brukes så er resultatet av målinger av bromkonsentrasjonen uavhengige og normalfordelte med kjent standardavvik $\sigma = 20$ (ppm).

La $X$ betegne bromkonsentrasjonen i en tilfeldig måling.

a) Dersom $\mu = 175$, regn ut $P(X > 200)$ og $P(150 < X < 200)$.
Hvilken verdi må $\mu$ har dersom vi skal ha at $P(X > 200) = 0.25$?

Bromkonsentrasjonen i produksjonsvannet fra brønn X10 i utkanten av et oljereservoar har lenge ligget på et nivå med $\mu = 175$, mens brønner i andre deler av reservoaret har hatt høyere nivåer. Etter at en ny brønn, Y16, som ligger i nærheten av X10 er satt i produksjon regner man med at flyten i reservoaret vil endre seg og at man bl.a. vil se det ved at bromkonsentrasjonen fra X10 vil gå opp.

Resultatet av seks målinger av bromkonsentrasjon fra brønn X10 etter åpning av den nye brønnen gav resultatene:

| | | | | | |
|---|---|---|---|---|---|
| 186 | 193 | 152 | 169 | 201 | 197 |

b) Gir dataene over grunnlag for å konkludere at forventet bromkonsentrasjon er høyere enn 175? Formuler problemstillingen som en hypotesetest, utfør testen og forklar hva resultatet betyr i praksis. Bruk 5% nivå.
Regn ut testens $p$-verdi.

I andre deler av reservoaret ligger forventet bromkonsentrasjon på rundt 190.

c) Regn ut hvor stor styrke testen formulert i punkt b) har for alternativet $\mu = 190$ når den utføres med seks målinger.
Hvor mange målinger måtte vært utført for at testen i punkt b) skulle hatt en styrke på minst 90% for alternativet $\mu = 190$?

Man er interessert i å sammenligne nivået til bromkonsentrasjonen i de to brønnene X10 og Y16. La $\mu_X$ betegne forventet bromkonsentrasjon i X10 og la $\mu_Y$ betegne forventet bromkonsentrasjon i Y16. Videre er $\overline{X}$ gjennomsnittet av $n_X$ målinger fra X10 og $\overline{Y}$ er gjennomsnittet av $n_Y$ målinger fra Y16. Som før antas alle konsentrasjonsmålinger å være uavhengige og normalfordelte med $\sigma = 20$.

En rimelig estimator for differansen i forventet bromkonsentrasjon mellom de to brønnene, $\mu_X - \mu_Y$, er gitt ved: $\hat{\mu}_X - \hat{\mu}_Y = \overline{X} - \overline{Y}$.

d) Vis at $\hat{\mu}_X - \hat{\mu}_Y$ er forventningsrett og at $\text{Var}(\hat{\mu}_X - \hat{\mu}_Y) = \sigma^2(1/n_X + 1/n_Y)$.
Bruk resultatene over som utgangspunkt og finn et 95% konfidensintervall for $\mu_X - \mu_Y$. Regn ut intervallet når det er gjort 10 målinger fra hver brønn og $\overline{x} = 186$ og $\overline{y} = 197$.
Kan vi fra konfidensintervallet konkludere med at det er forskjell i forventet bromkonsentrasjon mellom de to brønnene? Forklar kort.