Løsning eksamen 14. mai 2018

Oppgave 1:

a) Vi har en situasjon karakterisert ved:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" - flere deler som er har feil eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, p - samme sannsynlighet $p = 0.08$ for feil for hver del.
- Enkeltforsøkene er uavhengige - uavhengig fra del til del om de har feil.
- Et bestemt antall, n enkeltforsøk - et bestemt antall deler n som er valgt.

Dermed er X="antall deler som har indre feil" binomisk fordelt med parametre n og $p = 0.08$.

Med $X \sim \text{Bin}(20, 0.08)$ får vi:

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - (\binom{20}{0}(0.08)^0(1-0.08)^{20-0} + \binom{20}{1}(0.08)^1(1-0.08)^{20-1}) \\ &= 1 - (0.1887 + 0.3282) = \underline{\underline{0.48}} \end{aligned}$$

Siden $np(1-p) = 200 \cdot 0.08 \cdot (1 - 0.08) = 14.7 > 5$ kan vi bruke tilnærming til normalfordeling:

$$\begin{aligned} P(X \geq 20) &= 1 - P(X \leq 19) \approx 1 - P(Z \leq \frac{19 + 0.5 - \text{E}(X)}{\sqrt{\text{Var}(X)}}) = 1 - P(Z \leq \frac{19 + 0.5 - np}{\sqrt{np(1-p)}}) \\ &= 1 - P(Z \leq \frac{19 + 0.5 - 200 \cdot 0.08}{\sqrt{200 \cdot 0.08 \cdot 0.92}}) = 1 - P(Z \leq 0.91) = 1 - 0.8186 = \underline{\underline{0.18}} \end{aligned}$$

(Om man utelater heltallskorreksjonen +0.5, får man enten svaret 0.22 dersom man starer ut som over, eller svaret 0.15 dersom man starter ut med $P(X \geq 20) = 1 - P(X < 20)$).

Merk at siden $n > 10$ og $p < 0.1$ kunne man også brukt tilnærming til Poisson-fordeling men dette ville her ikke forenklet utregningene.

b) $P(F) = \underline{\underline{0.08}}$, $P(U|F) = \underline{\underline{0.96}}$ og $P(U|\overline{F}) = \underline{\underline{0.06}}$

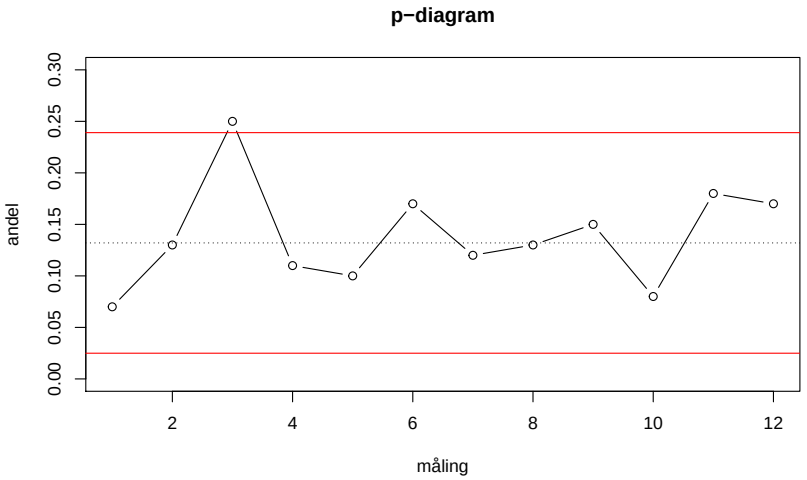
$$\begin{aligned} P(\overline{F}) &= 1 - P(F) = 1 - 0.08 = \underline{\underline{0.92}} \\ P(\overline{U}|\overline{F}) &= 1 - P(U|\overline{F}) = 1 - 0.06 = \underline{\underline{0.94}} \\ P(U) &= P(U \cap F) + P(U \cap \overline{F}) = P(U|F)P(F) + P(U|\overline{F})P(\overline{F}) \\ &= 0.96 \cdot 0.08 + 0.06 \cdot 0.92 = \underline{\underline{0.132}} \\ P(F|\overline{U}) &\overset{Bayes}{=} \frac{P(\overline{U}|F)P(F)}{P(\overline{U})} = \frac{(1 - P(U|F))P(F)}{1 - P(U)} = \frac{(1 - 0.96) \cdot 0.08}{1 - 0.132} = \underline{\underline{0.004}} \end{aligned}$$

c) Merk at siden vi i denne situasjonen kjenner p trenger vi ikke estimere p fra dataene. Senterlinja i diagrammet blir da bare $p = 0.132$ og for kontrollgrensene får vi:

Øvre kontrollgrense: $p + 3\sqrt{p(1-p)/n} = 0.132 + 3\sqrt{0.132(1-0.132)/100} = \underline{\underline{0.234}}$.

Nedre kontrollgrense: $p - 3\sqrt{p(1-p)/n} = 0.132 - 3\sqrt{0.132(1-0.132)/100} = \underline{\underline{0.030}}$.

Plott av $\hat{p}$-diagrammet er gitt under:



Vi ser at ved den tredje registreringen er prosessen utenfor kontrollgrensene - i denne perioden har det trolig skjedd noe som har medført unormalt høy andel produserte delere med indikasjon på indre feil.

Den praktiske tolkningen av en alarm over øvre kontrollgrenser er at enten har det skjedde noe i produksjonsprosessen slik at andel deler med feil har økt, eller så har det skjedd noe med ultralydtesten slik at andel "falske positive" (andel deler som er ok men der testen feilaktig slår ut) har økt.

Oppgave 2:

a) La Y være antall jordskjelv i løpet av ett år. Fra opplysningene i oppgaven vil da Y være Poisson-fordelt med $\lambda = 10$ og $t = 1$.

$$P(Y = 8) = \frac{10^8}{8!}e^{-10} = \underline{\underline{0.11}}$$

La M være antall jordskjelv i løpet av et halvt år. Da vil M være Poissonfordelt med $\lambda = 10$ og $t = 0.5$, dvs forventning $\lambda t = 10 \cdot 0.5 = 5$.

$$P(M > 4) = 1 - P(Y \leq 4) \overset{\text{tabell}}{=} 1 - 0.440 = \underline{\underline{0.56}}$$

Evt kan man summere sannsynlighetene for alle muligheter fra 0 til 4 i stedet for å bruke tabellen.

La T være tiden mellom to etterfølgende skjelv. Da er T eksponentialfordelt med $\lambda = 10$. Husk også at to måneder er $1/6$ år.

$$P(T > 1/6) = 1 - P(T \leq 1/6) = 1 - (1 - e^{-10 \cdot (1/6)}) = e^{-10/6} = \underline{\underline{0.19}}$$

Eventuelt kan man løse oppgaven med å bruke at dersom det skal gå mer enn to måneder mellom to jordskjelv så betyr dette at det kommer null jordskjelv i løpet av en periode på to måneder. Dersom X er antall jordskjelv i løpet av to måneder er X Poisson-fordelt med $\lambda = 10$ og $t = 1/6$ og:

$$P(X = 0) = \frac{(10/6)^0}{0!}e^{-10/6} = e^{-10/6} = \underline{\underline{0.19}}$$

Oppgave 3:

a) Den generelle lineære regresjonsmodellen er $Y = \alpha + \beta x + e$ der vi antar at $e \sim N(0, \sigma)$ og vi antar at feilleddene $e_1, \ldots, e_n$ for ulike målinger er uavhengige.

Vi har et årsaks/virknings -forhold her som er slik at mengden tilsetningsstoff påvirker hardheten (og ikke motsatt), og da er det rimelig å sette tilsetningsstoff som x-variabel og hardhet som Y-variabel. Det er også slik i denne situasjonen at mengden tilsetning er en variabel man selv bestemmer, dvs den er ikke-stokastisk, mens den resulterende hardheten er stokastisk. I slike situasjoner settes den ikke-stokastiske variabelen som x-variabel og den stokastiske som Y-variabel.

Residualet til en observasjon er definert som $\underline{\underline{\epsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i}}$ og for første observasjon blir dette $\epsilon_1 = y_1 - 35.8 - 2.49 \cdot x_1 = 37 - 35.8 - 2.49 \cdot 4 = \underline{\underline{-8.76}}$.

Residualplottet er et plott av residualene mot x (mengde tilsetningsstoff) og fra dette plottet kan vi sjekke om antagelsen om linær sammenheng mellom x og forventningen til Y holder, og om antagelsen om lik varians i Y for alle x-verdier holder. Dersom disse antagelsen holder skal vi ikke se noe bestemt mønster i dette plottet (residualene skal være symmetrisk fordelt om 0 og med lik varians for alle x-verdier). Plottet viser noenlunde lik varians for alle x-verdier, men vi ser et klart mønster i residualplottet. I starten ligger alle residualene under 0, så over og så til slutt under igjen. Dette er en sterk indikasjon på at antagelsen om lineær sammenheng ikke holder. Det ser vi også på plott av dataene med regresjonslinjen tegnet inn, dataene viser en klar ikke-lineær tendens og passer dårlig overens med den rette linja. Dvs den lineære regresjonsmodellen er ikke en god modell for å beskrive disse dataene.

b) Estimert kurve: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 = \underline{\underline{-36.4 + 22.9 \cdot x - 1.32 \cdot x^2}}$.

Residualet blir nå: $\epsilon_1 = y_1 - (-36.4 + 22.9 \cdot x_1 - 1.32 \cdot x_1^2) = 37 - (-36.4 + 22.9 \cdot 4 - 1.32 \cdot 4^2) = \underline{\underline{2.92}}$.

Både ut fra plottet av dataene med regresjonskurven tegnet inn og ut fra plottet av residualene så ser dette ut til å være en mye bedre modell. I plottet av dataene så passer andregradskurven bedre overens med dataeene enn den rette linjen i den første modellen. I residualplottet så er residualene jevnt fordelt rundt null uten noe særlig mønster.

Vi kan finne toppunktet til regresjonskurven ved å derivere og sette lik null:

$$\frac{d\hat{y}}{dx} = 22.9 - 2 \cdot 1.32x = 0 \quad \Rightarrow \; x = 22.9/(2 \cdot 1.32) = 8.7$$

Dvs i følge den estimerte modellen er hardheten høyest dersom mengden tilsetningsstoff er $\underline{\underline{8.7}}$.

Vi trenger andregradsleddet i modellen dersom $\beta_2 \neq 0$, dvs vi skal teste:

$$H_0: \;\; \beta_2 = 0 \qquad \text{mot} \qquad H_1: \;\; \beta_2 \neq 0$$

Vi leser p-verdien for testen rett ut fra datautskriften. Vi ser at for andregradsleddet i modellen så er p-verdien$=1.02 \cdot 10^{-7} < 0.05$ dvs $\underline{\underline{\text{vi forkaster } H_0}}$ og konkluderer med at andregradsleddet trengs.

Oppgave 4:

a)

$$\begin{aligned}
P(X > 200) &= 1 - P(\frac{X - 175}{20} < \frac{200 - 175}{20}) = 1 - P(Z < 1.25) = 1 - 0.8944 = \underline{\underline{0.11}} \\
P(150 < X < 200) &= P(X < 200) - P(X < 150) = P(Z < \frac{200 - 175}{20}) - P(Z < \frac{150 - 175}{20}) \\
&= P(Z < 1.25) - P(Z < -1.25) = 0.8944 - 0.1056 = \underline{\underline{0.79}} \\
P(X > 200) &= 0.25 \;\Rightarrow\; P(X < 200) = 0.75 \\
P(Z < \frac{200 - \mu}{20}) &= 0.75 \;\Rightarrow\; \frac{200 - \mu}{20} = 0.67 \\
& \quad \mu = -0.67 \cdot 20 + 200 = \underline{\underline{187}}
\end{aligned}$$

b) Vi har uavhengige normalfordelte målinger med kjent standardavvik $\sigma = 20$ og skal teste:

$$H_0: \ \mu \leq 175 \qquad \text{mot} \qquad H_1: \ \mu > 175$$

Estimator: $\hat{\mu} = \bar{X}$. Siden σ er kjent har vi dersom H_0 er korrekt følgende nullfordeling:

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{Var}(\overline{X})}} = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - 175}{20/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Med signifikansnivå $\alpha = 0.05$ forkaster vi H_0 dersom $Z \geq z_{0.05} = 1.645$. Gjennomsnittet av målingene er $\overline{x} = (186 + 193 + 152 + 169 + 201 + 197)/6 = 183$ og observert verdi på testobservatoren blir da:

$$z_{obs} = \frac{183 - 175}{20/\sqrt{6}} = 0.98 < 1.645.$$

Dvs. utfallet er i akseptsområdet. Konklusjon: Vi $\underline{\underline{\text{forkaster ikke } H_0}}$. Vi kan ikke konkludere at forventet bromkonsentrasjon er over 175.

$$p - \text{verdi} = P(Z > z_{obs}) = P(Z > 0.98) = 1 - P(Z < 0.98) = 1 - 0.8365 = \underline{\underline{0.16}}$$

c) Merk at vi har har en test av typen hvor $H_1: \ \mu > \mu_0$, og vi har da fra formelarket at vi skal bruke $\gamma(\mu) = 1 - P(Z \leq z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}})$. Med $\sigma = 20$, $n = 6$, $z_\alpha = z_{0.05} = 1.645$ og $\mu_0 = 175$ blir beregningen:

$$\gamma(190) = 1 - P(Z \leq 1.645 + \frac{175 - 190}{20/\sqrt{6}}) = 1 - P(Z \leq -0.19) = 1 - 0.4247 = \underline{\underline{0.58}}$$

Dette betyr i praksis at dersom $\mu = 190$ er det 58% sannsynlighet for at testen i punkt b) vil gi forkastning når $n = 6$.

En styrke på 90% betyr at $1 - \beta = 0.90$ eller $\beta = 0.10$ (der β =P(type II feil)). Da er $z_\beta = z_{0.10} = 1.282$. Formelen på formelarket gir oss da at nødvendig utvalgsstørrelser blir

$$n = \frac{(z_\beta + z_\alpha)^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(z_{0.1} + z_{0.05})^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(1.282 + 1.645)^2 20^2}{(175 - 190)^2} = 15.2$$

Dvs de må gjøres minst $\underline{\underline{16}}$ målinger for å få styrke på minst 0.90 (som er det samme som en sannsynlighet for type II feil på maks 0.10).

d) Husk først at med $\overline{X} = \frac{1}{n_X}\sum_{i=1}^{n_X} X_i$ der $X_1, \ldots X_{n_X}$ er uavhengige med $\mathrm{E}(X) = \mu$ og $\mathrm{Var}(X) = \sigma^2$ så vil $\mathrm{E}(\overline{X}) = \frac{1}{n_X}\sum_{i=1}^{n_X}\mathrm{E}(X_i) = \frac{1}{n_X}\sum_{i=1}^{n_X}\mu_X = \mu_X$ og $\mathrm{Var}(\overline{X}) = \frac{1}{n_X^2}\sum_{i=1}^{n_X}\mathrm{Var}(X_i) = \frac{1}{n_X^2}\sum_{i=1}^{n_X}\sigma^2 = \sigma^2/n_X$. (Dette står også nederst på side 2 av formelarkene og kan brukes som kjent.)

Tilsvarende vil $\mathrm{E}(\bar{Y}) = \mu_Y$ og $\mathrm{Var}(\bar{X}) = \sigma^2/n_Y$, og vi får da:

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_X - \hat{\mu}_Y) &= \mathrm{E}(\overline{X} - \overline{Y}) = \mathrm{E}(\overline{X}) - \mathrm{E}(\overline{Y}) = \underline{\underline{\mu_X - \mu_Y}} \\
\mathrm{Var}(\hat{\mu}_X - \hat{\mu}_Y) &= \mathrm{Var}(\overline{X} - \overline{Y}) = \mathrm{Var}(\overline{X}) + \mathrm{Var}(\overline{Y}) \\
&= \sigma^2/n_X + \sigma^2/n_Y = \underline{\underline{\sigma^2(1/n_X + 1/n_Y)}}
\end{aligned}$$

Siden estimatoren $\hat{\mu}_X - \hat{\mu}_Y = \overline{X} - \overline{Y}$ er en lineærkombinasjon av uavhengige normalfordelte variabler er den normalfordelt og vi har dermed at

$$P(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2(1/n_X + 1/n_Y)}} < z_{\alpha/2}) = 1 - \alpha$$

som gir konfidensintervallet:

$$[\overline{X} - \overline{Y} - z_{\alpha/2}\sigma\sqrt{1/n_X + 1/n_Y}\ ,\ \overline{X} - \overline{Y} + z_{\alpha/2}\sigma\sqrt{1/n_X + 1/n_Y}]$$

For et 95% intervall setter vi $\alpha = 0.05$ som gir $z_{\alpha/2} = z_{0.025} = 1.96$. Innsatt observerte data får vi:

$$[186 - 197 - 1.96 \cdot 20\sqrt{1/10 + 1/10}\ ,\ 186 - 197 + 1.96 \cdot 20\sqrt{1/10 + 1/10}] = \underline{\underline{[-28.5\ ,\ 6.5]}}$$

Siden konfidensintervallet inneholder 0 kan vi ikke konkludere med at det er forskjell i forventet bromkonsentrasjon mellom de to brønnene. En test av nullhypotesen $\mu_X - \mu_Y = 0$ mot alternativet $\mu_X - \mu_Y \neq 0$ vil her ikke gi forkastning å 5% nivå siden 0 er inneholdt i 95% konfidensintervallet.