Løsning eksamen 14. mai 2019

Oppgave 1

a)

$$\begin{aligned}
\mathrm{E}(X) &= \sum_x xP(X=x) = 0\cdot 0.05 + 1\cdot 0.15 + 2\cdot 0.20 + 3\cdot 0.25 + 4\cdot 0.35 = \underline{\underline{2.7}} \\
\mathrm{E}(X^2) &= \sum_x x^2P(X=x) = 0^2\cdot 0.05 + 1^2\cdot 0.15 + 2^2\cdot 0.20 + 3^2\cdot 0.25 + 4^2\cdot 0.35 = 8.8 \\
\Rightarrow \mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 8.8 - 2.7^2 = \underline{\underline{1.51}} \\
Y &= 750X - 1150 \\
\mathrm{E}(Y) &= \mathrm{E}(750X - 1150) = 750\mathrm{E}(X) - 1150 = 750\cdot 2.7 - 1150 = \underline{\underline{875}} \\
\mathrm{Var}(Y) &= \mathrm{Var}(750X - 1150) = 750^2\mathrm{Var}(X) = 750^2\cdot 1.51 = \underline{\underline{849375}}
\end{aligned}$$

b)

$$\begin{aligned}
\mathrm{E}(W) &= \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \ldots \mathrm{E}(Y_{300}) = 300\cdot 875 = \underline{\underline{262500}} \\
\mathrm{Var}(W) &\overset{\text{uavh.}}{=} \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \ldots \mathrm{Var}(Y_{300}) = 300\cdot 849375 = 254812500 \\
\mathrm{SD}(W) &= \sqrt{\mathrm{Var}(W)} = \sqrt{254812500} = \underline{\underline{15962.85}}
\end{aligned}$$

$\mathrm{E}(W)$ er forventet årlig fortjeneste.

SGT gir at en sum av mange uavhengige variable er tilnærmet normalfordelt
$\Rightarrow W = Y_1 + \cdots + Y_{300} \approx N(\mathrm{E}(W), \mathrm{SD}(W)) = N(262500, 15962.85)$

$$\begin{aligned}
P(W > 250000) &= 1 - P(W \leq 250000) \approx 1 - P(Z \leq \frac{250000 - 262500}{15962.85}) \\
&= 1 - P(Z \leq -0.78) = 1 - 0.2177 = \underline{\underline{0.78}}
\end{aligned}$$

Oppgave 2

a)

$$\begin{aligned}
P(X < 0.15) &= P(\frac{X - 0.18}{0.02} < \frac{0.15 - 0.18}{0.02}) = P(Z < -1.50) = \underline{\underline{0.0668}} \\
P(0.15 < X < 0.20) &= P(X < 0.20) - P(X < 0.15) \\
&= P(\frac{X - 0.18}{0.02} < \frac{0.20 - 0.18}{0.02}) - P(\frac{X - 0.18}{0.02} < \frac{0.15 - 0.18}{0.02}) \\
&= P(Z < 1.00) - P(Z < -1.50) = 0.8413 - 0.0668 = \underline{\underline{0.7745}} \\
P(X < k) &= P(Z < \frac{k - 0.18}{0.02}) = 0.99 \;\Rightarrow\; \frac{k - 0.18}{0.02} = 2.33 \\
&\quad k = 2.33\cdot 0.02 + 0.18 = \underline{\underline{0.2266}}
\end{aligned}$$

Husk at $\mathrm{E}(\bar{X}) = \mu$ og $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

$$\begin{aligned}
P(\bar{X} < 0.15) &= P(\frac{\bar{X} - \mathrm{E}(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}} < \frac{0.15 - \mathrm{E}(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}}) = P(Z < \frac{0.15 - 0.18}{\sqrt{\frac{0.02^2}{5}}}) \\
&= P(Z < -3.35) = \underline{\underline{0.0004}}
\end{aligned}$$

Sannsynligheten for at gjennomsnittet av 5 uavhengige målinger er mindre enn 0.15 er bare 0.004, og dette er mye mindre enn sannsynligheten for at en tilfeldig måling er mindre enn 0.15 som fra første spørsmål er 0.0668. Et gjennomsnittet har mindre varians enn enkeltmålinger og det er derfor mindre sannsynliget for at et gjennomsnitt skal falle så langt fra forventningsverdien enn at en enkeltmåling skal gjøre det.

b) $(1-\alpha)100\%$ konfidensintervall for μ når σ kjent er gitt ved: $[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$.
Med $\alpha = 0.01$ blir $z_{\alpha/2} = z_{0.005} = 2.575$, og med $\bar{x} = 0.142$, $n = 12$ og $\sigma = 0.02$ blir 99% konfidensintervallet:

$$[0.142 - 2.575\frac{0.02}{\sqrt{12}}, 0.142 + 2.575\frac{0.02}{\sqrt{12}}] = \underline{\underline{[0.127, 0.157]}}$$

Lengden til konfidensintervallet er:

$$\begin{aligned} L &= \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - (\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \text{Dvs: } L \le 0.01 \quad &\Rightarrow \quad 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le 0.01 \quad \Rightarrow \quad 200z_{\alpha/2}\sigma \le \sqrt{n} \\ & \qquad n \ge (200z_{\alpha/2}\sigma)^2 = (200 \cdot 2.575 \cdot 0.02)^2 = 10.3^2 = 106.09 \end{aligned}$$

Dvs det må gjøres minst $\underline{\underline{107}}$ målinger for å få et konfidensintervall med ønsket lengde.

c) $\qquad H_0: \ \mu \ge 0.15 \qquad$ mot $\qquad H_1: \ \mu < 0.15$

Situasjonen her er normalfordeling med ukjent μ og kjent σ. Dersom H_0 er korrekt er da

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, forkaster vi H_0 dersom $Z \le -z_{0.05} = -1.645$.
Observert: $z = \frac{0.142 - 0.15}{0.02/\sqrt{12}} = -1.39$
Siden $-1.39 > -1.645$ blir konklusjonen at vi $\underline{\underline{\text{ikke forkaster } H_0}}$. Dataene gir ikke grunnlag for å konkludere, på 5% nivå, at forventet mengde fosfor er mindre enn 0.15.

p-verdien blir her (siden vi har ensidig test av type "mindre enn"):

$$p - \text{verdi} = P(Z \le z_{obs}) = P(Z < -1.39) = \underline{\underline{0.0823}}$$

Dvs vi forkaster ikke testen på 5% nivå.

d) Styrkefunksjonen gir oss sannsynligheten for å forkaste H_0 for ulike verdier av μ.

For beregning av styrke så bruker vi formelen på formelarket (eller side 29/33 i notatene om hypotesetesting, eller regel 6.16 i boka). Merk at vi har har en test av typen hvor $H_1: \ \mu < \mu_0$, og vi skal da bruke $\gamma(\mu) = P(Z \le -z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}})$. Med $\sigma = 0.02$, $n = 12$, $z_\alpha = z_{0.05} = 1.645$ og $\mu_0 = 0.15$ blir beregningen:

$$\begin{aligned} \gamma(0.14) &= P(Z \le -1.645 + \frac{0.15 - 0.14}{0.02/\sqrt{12}}) = P(Z \le 0.09) = \underline{\underline{0.54}} \\ \gamma(0.13) &= P(Z \le -1.645 + \frac{0.15 - 0.13}{0.02/\sqrt{12}}) = P(Z \le 1.82) = \underline{\underline{0.966}} \end{aligned}$$

En styrke på 95% betyr at $1 - \beta = 0.95$ eller $\beta = 0.05$ (der β =P(type II feil)). Da er $z_\beta = z_{0.05} = 1.645$. Formelen på formelarket/forelesningsnotatene gir oss da at nødvendig utvalgsstørrelser blir

$$n = \frac{(z_\beta + z_\alpha)^2\sigma^2}{(\mu_0 - \mu)^2} = \frac{(z_{0.05} + z_{0.05})^2\sigma^2}{(0.15 - 0.14)^2} = \frac{(1.645 + 1.645)^2 0.02^2}{(0.15 - 0.14)^2} = 43.3$$

Dvs de må gjøres minst $\underline{\underline{44}}$ målinger for å få styrke på minst 0.95 (som er det samme som en sannsynlighet for type II feil på maks 0.05).

Oppgave 3:

a) Estimert regresjonsline: $\hat{y} = \hat{\alpha} + \hat{\beta}x = \underline{\underline{66.5 - 0.34 \cdot x}}$.
År 2020 tilsvarer $x = 50$ og gir $\hat{y} = 66.5 - 0.34 \cdot 50 = \underline{\underline{49.5}}$.

Fra det første residualplottet kan man sjekke om antagelsen om lineære sammenheng mellom x og forventningen til Y holder, og om antagelsen om lik varians i Y for alle x-verdier holder. Videre kan vi i denne spesielle situasjonen fra dette plottet også sjekket om der er avhengigheter over tid siden vi faktisk har innsamlingsrekkefølgen som x-variabel. Alle disse antagelsene ser ut til å være oppfyllte siden vi ikke har noe bestemt mønster i residualene. Det andre residualplottet er en sjekk av om antagelsen om normalfordelt feilledd er oppfylt, og det ser også ok ut siden punktene her faller greit på den rette linjen. Dvs, antagelsene for regresjonsmodellen ser ut for å være oppfyllte.

Dersom $\beta \neq 0$ i regresjonsmoodellen betyr det at forventet lengde på issesongen har endret seg over tid, dvs vi skal teste:

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

Vi leser p-verdien for testen rett ut fra datautskriften. Vi ser a p-verdien=0.0097 < 0.05 dvs $\underline{\underline{\text{vi forkaster } H_0}}$ og konkluderer med at forventet lengde på issesongen har endret seg over perioden vi har data fra. Siden den estimerte verdien på β er negativ kan vi konkludere at forventet lengde på issesongen er redusert.

Estimert forventet lengde på issesongen i 1971: $\hat{y} = 66.5 - 0.34 \cdot 1 = 66.16$. Estimert forventet lengde på issesongen i 2019: $\hat{y} = 66.5 - 0.34 \cdot 49 = 49.84$. Estimert endring: 66.16-49.84=16.32, dvs en reduksjon på $\underline{\underline{16.3}}$ dager i forventet lengde på issesongen.

Oppgave 4:

Siden vi har for eksponentialfordeling at $\mathrm{E}(T) = 1/\lambda$ får vi her at $1/\lambda = 2.5$ dvs $\lambda = 0.4$

a) Vi kan finne sannsynligheten enten ved å integrere sannsynlighetstettheten, eller, litt enklere, ved å bruke den kumulative fordelingsfunksjonen: $P(T < t) = F(t) = 1 - e^{-\lambda t} = 1 - e^{-0.4t}$. Vi får da:

$$\begin{aligned}
P(T < 1) &= F(1) = 1 - e^{-0.4 \cdot 1} = \underline{\underline{0.33}} \\
P(1 < T < 2) &= P(T < 2) - P(T < 1) = F(2) - F(1) \\
&= 1 - e^{-0.4 \cdot 2} - (1 - e^{-0.4 \cdot 1}) = \underline{\underline{0.22}} \\
P(T < t_{0.5}) &= F(t_{0.5}) = 1 - e^{-0.4 \cdot t_{0.5}} = 0.5 \\
0.5 &= e^{-0.4 \cdot t_{0.5}} \ \Rightarrow \ t_{0.5} = \ln(0.5)/(-0.4) = \underline{\underline{1.73}}
\end{aligned}$$

b) Vi har en situasjon karakterisert ved:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" - flere lysrør som feiler innen ett år eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, p - samme sannsynlighet $p = 0.33$ for feil for hvert lysrør.
- Enkeltforsøkene er uavhengige - uavhengig fra rør til rør om de har feil.
- Et bestemt antall, n enkeltforsøk - et bestemt antall lysrør n som installeres.

Dermed er X="antall lysrør som feiler innen ett år" binomisk fordelt med parametre n og $p = 0.33$.

Med $X \sim \text{Bin}(10, 0.33)$ får vi:

$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) = 1 - (P(X=0) + P(X=1) + P(X=2) + P(X=3)) \\ &= 1 - \Big(\binom{10}{0}(0.33)^0(1-0.33)^{10-0} + \binom{10}{1}(0.33)^1(1-0.33)^{10-1} \\ &\qquad + \binom{10}{2}(0.33)^2(1-0.33)^{10-2} + \binom{10}{3}(0.33)^3(1-0.33)^{10-3} \Big) \\ &= 1 - (0.0182 + 0.0898 + 0.1990 + 0.2614) = \underline{\underline{0.43}} \end{aligned}$$

Siden $np(1-p) = 100 \cdot 0.33 \cdot (1 - 0.33) = 22.1 > 5$ kan vi bruke tilnærming til normalfordeling:

$$\begin{aligned} P(X > 30) &= 1 - P(X \leq 30) \approx 1 - P(Z \leq \frac{30 + 0.5 - \text{E}(X)}{\sqrt{\text{Var}(X)}}) = 1 - P(Z \leq \frac{30 + 0.5 - np}{\sqrt{np(1-p)}}) \\ &= 1 - P(Z \leq \frac{30 + 0.5 - 100 \cdot 0.33}{\sqrt{100 \cdot 0.33 \cdot 0.67}}) = 1 - P(Z \leq -0.53) = 1 - 0.2981 = \underline{\underline{0.70}} \end{aligned}$$

(Om man utelater heltallskorreksjonen +0.5, får man enten svaret 0.74 dersom man starer ut som over, eller svaret 0.67 dersom man starter ut med $P(X > 30) = P(X \geq 31) = 1 - P(X < 31)$).

c)

$$\begin{aligned} \text{E}(\hat{\beta}) &= \text{E}(\frac{1}{n}\sum_{i=1}^{n} T_i) = \frac{1}{n}\sum_{i=1}^{n} \text{E}(T_i) = \frac{1}{n}\sum_{i=1}^{n} \beta = \frac{1}{n} n\beta = \underline{\beta} \\ \text{Var}(\hat{\beta}) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} T_i) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(T_i) = \frac{1}{n^2}\sum_{i=1}^{n} \beta^2 = \frac{1}{n^2} n\beta^2 = \underline{\underline{\frac{\beta^2}{n}}} \end{aligned}$$

Siden estimatoren $\hat{\beta} = \overline{T}$ er et gjennomsnitt av mer enn 30 uavhengige stokastiske variabler fra samme fordeling følger det fra sentralgrenseteoremet at estimatoren er tilnærmet normalfordelt. Vi har dermed at

$$P(-z_{\alpha/2} < \frac{\hat{\beta} - \beta}{\sqrt{\beta^2/n}} < z_{\alpha/2}) \approx P(-z_{\alpha/2} < \frac{\hat{\beta} - \beta}{\sqrt{\hat{\beta}^2/n}} < z_{\alpha/2}) = 1 - \alpha$$

som gir konfidensintervallet:

$$[\hat{\beta} - z_{\alpha/2}\sqrt{\hat{\beta}^2/n}\ ,\ \hat{\beta} + z_{\alpha/2}\sqrt{\hat{\beta}^2/n}]$$

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

$$[2.38 - 1.96 \cdot \sqrt{2.38^2/50}\ ,\ 2.38 + 1.96 \cdot \sqrt{2.38^2/50}] = \underline{\underline{[1.7\ ,\ 3.0]}}$$