Løsning prøveeksamen 1. april 2020

Oppgave 1

a)

$$\begin{aligned} P(X > 100) &= 1 - P(\frac{X-90}{10} \leq \frac{100-90}{10}) = 1 - P(Z \leq 1) = 1 - 0.8413 = \underline{\underline{0.1587}} \\ P(X > a) = 0.01 &\Rightarrow P(X \leq a) = P(Z \leq \frac{a-90}{10}) = 0.99 \Rightarrow \frac{a-90}{10} = 2.33 \\ &\quad a = 2.33 \cdot 10 + 90 = \underline{\underline{113.3}} \end{aligned}$$

Siden $\mathrm{E}(\bar{X}) = \mu = 90$ og $\mathrm{Var}(\bar{X}) = \sigma^2/n = 10^2/5 = 20$ får vi:

$$P(\bar{X} > 100) = 1 - P(\frac{\bar{X}-90}{\sqrt{20}} \leq \frac{100-90}{\sqrt{20}}) = 1 - P(Z \leq 2.24) = 1 - 0.9875 = \underline{\underline{0.0125}}$$

b) La X_A være lengden av en laks fra stamme A og X_B være lengden av en laks fra stamme B. Vi har i A regnet ut at $P(X_A > 100) = 0.1587$. For en laks fra stamme B får vi tilsvarende:

$$P(X_B > 100) = 1 - P(\frac{X_B - 96}{8} \leq \frac{100-96}{8}) = 1 - P(Z \leq 0.5) = 1 - 0.6915 = \underline{\underline{0.3085}}$$

Setningen om total sannsynlighet gir da at:

$$P(X > 100) = P(X > 100|A)P(A) + P(X > 100|B)P(B) = 0.1587 \cdot 0.70 + 0.3085 \cdot 0.30 = \underline{\underline{0.204}}$$

Merk at $X_B < X_A$ er det samme som at $X_B - X_A < 0$ og vi får da:

$$\begin{aligned} P(X_B < X_A) &= P(X_B - X_A < 0) = P(\frac{X_B - X_A - \mathrm{E}(X_B - X_A)}{\sqrt{\mathrm{Var}(X_B - X_A)}} < \frac{0 - \mathrm{E}(X_B - X_A)}{\sqrt{\mathrm{Var}(X_B - X_A)}}) \\ &= P(Z < \frac{0 - (\mathrm{E}(X_B) - \mathrm{E}(X_A))}{\sqrt{\mathrm{Var}(X_B) + (-1)^2\mathrm{Var}(X_A)}}) = P(Z < \frac{0 - (96-90)}{\sqrt{10^2 + 8^2}}) = P(Z < -0.47) = \underline{\underline{0.319}} \end{aligned}$$

Oppgave 2

a) Modell: $Y_{ij} = \mu_i + e_{ij}$, der e_{ij} uavh. $\mathrm{N}(0, \sigma)$. Her er Y_{ij} målt smerteterskel for person j, med hårfarge i, μ_i er forventa smerteterskel for personer med hårfarge i, e_{ij} er feilleddet (tilfeldig variasjon) og σ er standardavviket for målt smerteterskel for hver hårfarge.

Dvs, vi antar normalfordeling og samme varians i hver gruppe, og uavhengighet mellom målingene.

(Alternativ formulering av modellen: $Y_{ij} = \mu + \alpha_i + e_{ij}$, der e_{ij} uavh. $\mathrm{N}(0, \sigma)$ der μ er gjennomsnittlig smerteterskel og α_i er effekten av hårfarge i.)

$$H_0: \ \mu_1 = \mu_2 = \mu_3 = \mu_3 \qquad \text{mot} \qquad H_1: \ \text{minst en } \mu_i \text{ ulik de andre}$$

Vi har fra pensum/formelarket at vi baserer denne testen på at under nullhypotesen er

$$F = \frac{SS_G/(k-1)}{SS_E/(N-k)} \sim F(k-1, N-k)$$

og vi forkaster nullhypotesen dersom F blir stor. Fra den oppgitte variansanalysetabellen ser vi at p-verdien for testen er $0.0041 < 0.05$, dvs $\underline{\underline{\text{vi forkaster } H_0 \text{ på 5\% nivå}}}$. Dette betyr i praksis at det er forskjell i forventet smerterskel for personer med ulike hårfarger.

Fra plottet av residualene mot gruppe (hårfarge) kan vi sjekke om antagelsen om like varians i hver gruppe ser ok ut - dette ser ut til å være oppfylt. Fra normalplottet kan vi se om antagelsen om normalfordeling ser ut til å være oppfylt - punktene i normalplottet følger den rette linja ganske godt så dette ser også ok ut.

b) La μ_2 være forventet smerteterskel for de mørkeblonde og la μ_4 være forventet smerteterskel for de mørkebrune. Vi skal da teste:

$$H_0: \ \mu_2 = \mu_4 \qquad \text{mot} \qquad H_1: \ \mu_2 \neq \mu_4$$

Siden vi har en uparet sammenligning av to utvalg med ukjent varians baserer vi testen på

$$T = \frac{\bar{X}_2 - \bar{X}_4}{S_p\sqrt{\frac{1}{n_2} + \frac{1}{n_4}}} \sim t(n_2 + n_4 - 2)$$

Med nivå 5%, dvs $\alpha = 0.05$, tosidig test og $n_2 = n_4 = 5$ forkaster vi H_0 dersom $T \leq -t_{0.025,8} = -2.306$ eller dersom $T \geq t_{0.025,8} = 2.306$.

Fra oppgitt informasjon finner vi først s_p:

$$\begin{aligned} s_p^2 &= \frac{(n_2-1)s_{X_2}^2 + (n_4-1)s_{X_4}^2}{n_4 + n_4 - 2} = \frac{4 \cdot 9.28^2 + 4 \cdot 8.32^2}{5+5-2} = 77.67 \\ s_p &= \sqrt{77.67} = 8.813 \end{aligned}$$

Observert verdi på testobservatoren blir da: $t = \frac{51.2-37.4}{8.813\sqrt{\frac{1}{5}+\frac{1}{5}}} = 2.48$

Siden $2.48 > 2.306$ blir konklusjonen at vi $\underline{\underline{\text{forkaster } H_0}}$. Dataene gir grunnlag for å konkludere at det er forskjell i forventet smerteterskel mellom de med mørkeblondt og de med mørkebrunt hår. Siden $\bar{x}_2 > \bar{x}_4$ ser vi at effekten går i retning av at de med mørkeblondt hår har høyere smerteterskel enn de med mørkebrunt.

Antagelsene som ligger til grunn er at målingene er uavhengige og normalfordelte, med samme forventningsverdi i hver gruppe og med samme varians for alle målinger.

Oppgave 3

a) Vi kan finne sannsynligheten enten ved å integrere sannsynlighetstettheten, eller, litt enklere, ved å bruke den kumulative fordelingsfunksjonen: $P(T < t) = F(t) = 1 - e^{-\lambda t} = 1 - e^{-t/1.6} = 1 - e^{-0.625t}$. Her har vi brukt at i eksponentialfordelingen er $\mathrm{E}(T) = 1/\lambda$ og vi har oppgitt at $\mathrm{E}(T) = 1.6$ slik at $\lambda = 1/1.6$. Vi får da:

$$\begin{aligned} P(T > 1) &= 1 - F(1) = 1 - (1 - e^{-0.625 \cdot 1}) = e^{-0.625} = \underline{\underline{0.535}} \\ P(T < 1.6) &= 1 - e^{-0.625 \cdot 1.6} = 1 - e^{-1}) = \underline{\underline{0.632}} \end{aligned}$$

I siste spørsmål har en situasjon karakterisert ved:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" - flere lysrør som fungerer etter ett år eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, p - samme sannsynlighet $p = 0.535$ for å fungere for hvert lysrør.
- Enkeltforsøkene er uavhengige - uavhengig fra rør til rør om de fungerer.
- Et bestemt antall, n enkeltforsøk - et bestemt antall lysrør $n = 8$ som installeres.

Dermed er X="antall lysrør som fungerer etter ett år" binomisk fordelt med parametre $n = 8$ og $p = 0.535$.

Med $X \sim \text{Bin}(8, 0.535)$ får vi:

$$\begin{aligned} P(X \geq 6) &= P(X=6) + P(X=7) + P(X=8) \\ &= \binom{8}{6}(0.53)^6(1-0.535)^{8-6} + \binom{8}{7}(0.53)^7(1-0.535)^{8-7} + \binom{8}{8}(0.53)^8(1-0.535)^{8-8} \\ &= 0.1420 + 0.0467 + 0.0067 = \underline{\underline{0.195}} \end{aligned}$$

b) La Y være antall lysrør som feiler i løpet av en måned. Fra opplysningene i oppgaven vil da Y være Poissonfordelt med forventning $\lambda t = 45 \cdot (1/12) = 3.75$. Vi får da

$$\begin{aligned} P(Y \geq 3) &= 1 - P(Y < 3) = 1 - (P(X=0) + P(X=1) + P(X=2)) \\ &= 1 - \left(\frac{3.75^0}{0!}e^{-3.75} + \frac{3.75^1}{1!}e^{-3.75} + \frac{3.75^2}{2!}e^{-3.75} \right) = 1 - (0.0235 + 0.0882 + 0.1654) = \underline{\underline{0.72}} \end{aligned}$$

La X antall lysrør som feiler i løpet av ett år. Vi har da $t = 1$ og med $\lambda = 45$ har vi at X er Poissonfordelt med forventning $\lambda t = 45 \cdot 1 = 45$. Siden $\lambda t > 10$ kan vi bruke tilnærming til normalfordeling. Husk at i Poissonfordeling er $\text{E}(X) = \text{Var}(X) = \lambda t$.

$$P(X \geq 36) = 1 - P(X \leq 35) = 1 - P(Z \leq \frac{35 + 0.5 - 45}{\sqrt{45}} = 1 - P(Z \leq -1.42) = 1 - 0.0778 = \underline{\underline{0.92}}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da enten 0.93 eller 0.91 alt etter om man tar $P(X \geq 36) = 1 - P(X \leq 35)$ eller $P(X \geq 36) = 1 - P(X < 36)$ som utgangspunkt.)

c) Av samme grunn som forklart i siste spørsmål i punkt a) har vi også her at antall lysrør som fungerer etter ett år er binomisk fordelt. Estimat av p : $\hat{p} = 73/100 = \underline{\underline{0.73}}$.

Et (tilnærmet) $(1-\alpha)100\%$ konfidensintervall for p er gitt ved:

$$[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ , \ \ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Innsatt $\hat{p} = 0.73$, $n = 100$ og $z_{\alpha/2} = z_{0.025} = 1.96$ gir dette tilnærmet 95% konfidensintervall for p :

$$[0.73 - 1.96\sqrt{\frac{0.73(1-0.73)}{100}}, \ \ 0.73 + 1.96\sqrt{\frac{0.73(1-0.73)}{100}}] = \underline{\underline{[0.64 \ , \ \ 0.82]}}$$

$$p = P(T > 1) = \int_1^\infty (1/\beta)e^{-u/\beta}du = [-e^{-u/\beta}]_1^\infty = \underline{\underline{e^{-1/\beta}}}$$

Dersom vi starter fra siste skritt i utledningen av konfidensintervallet for p kan vi bruke relasjonen over og regne oss om til et konfidensintervall for β:

$$\begin{aligned} P(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) &= 1-\alpha \\ P(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < e^{-1/\beta} < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) &= 1-\alpha \\ P(\log\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) < -1/\beta < \log\left(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)) &= 1-\alpha \\ P(-1/\log\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) < \beta < -1/\log\left(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)) &= 1-\alpha \end{aligned}$$

Dvs intervallet blir

$$[-1/\log\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),\ \ -1/\log\left(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)]$$

og innsatt tall får vi:

$$[-1/\log(0.643),\ \ -1/\log(0.817)] = \underline{\underline{[2.26\ ,\ \ 4.95]}}$$

En ulempe med dette konfidensintervallet i forhold til et intervall basert på de eksakte tidene er at vi bruker mindre av informasjonen i dataene (vi bruker kun informasjonen om hvorvidt levetiden er større enn ett år eller ikke i stedet for å bruke informasjsonen om den eksakte levetiden). Dette medfører at vi får større usikkerhet og dermed et bredere konfidensintervall, dersom vi bruker samme antall målinger med de to fremgangsmåtene. Men vi kan kompansere for dette med å bruke flere målinger (registrere for et større antall lysrør om de fungerer i mer enn ett år eller ikke).

Oppgave 4

a) Vi har her en klar retning på årsak-virkning. Temperaturen påvirker isoleringsevnen og ikke motsatt. Videre så er også temperaturen ikke-stokastisk i dette eksperimentet mens isoleringsevnen er stokastisk. Begge disse forholdene tilsier at vi skal bruke temperatur som x-variabel og isoleringsevne som Y-variabel.

Fra datautskriften får vi at $\underline{\underline{\hat{y} = 30.5 - 0.083x}}$. Vi kan tolke det estimerte stigningstallet, -0.083, som at for hver grad temperaturen øker så reduseres isoleringsevnen med 0.083 kilovolt.

Prinsippet som brukes for å finne den estimerte regresjonslinja er å minimere den summerte kvadrat-avstanden mellom punkter og linja, kalles gjerne minste kvadraters metode. Dvs vi bruker som estimator de verdiene på α og β som minimerer $\sum_{i=1}^{n}(Y_i - (\alpha + \beta x_i))^2$.

Vi har at andel forklart variasjon er korrelasjon^2. Fra datautskriften har vi at $r^2 = 0.4913$. Dermed er korrelasjonen lik: $r = \pm\sqrt{0.4913}$. Siden det estimerte stigningstallet i modellen er negativt har vi en negativ korrelasjon mellom temperatur og isoleringsevne (jo høyere temperatur jo lavere isoleringsevne). Dermed blir: $r = -\sqrt{0.4913} = \underline{\underline{-0.70}}$.

b) Vi ser at det estimerte standardavviket er $s_1 = 2.983$ i modellen med bare temperatur, mens det er $s_2 = 2.287$ i modellen med både temperatur og tid. Dvs, det estimerte standardavviket er minst i den siste modellen. Dette er rimelig siden vi i denne modellen tar hensyn til effekten av både temperatur og tid, mens vi i den første modellen bare tar hensyn til effekten av tid. Vi ser i den siste modellen at tid har en høyt signifikant betydning og bidrar dermed til å forklare responsvariabelen (isoleringsevnen). I modellen som ikke tar med tid vil de da være større uforklart variasjon og dermed større standardavvik.

Vi ser også at R^2 og R^2-justert er klart høyere i modellen med både temperatur og tid (begge øker fra 0.49 til 0.70), og dette tyder på at den siste modellen er klart best. (Vi må se på R^2-justert når vi sammenligner modeller med ulike antall forklaringsvariabler da R^2 alltid vil øke når vi legger til flere variabler, og her er altså også R^2-justert klart høyere i den siste modellen).

Vi skal teste (merk at dersom reduksjonen er saktere enn 0.1 per uke betyr det at $\beta_2 > -0.1$):

$$H_0:\ \ \beta_2 \leq -0.1 \qquad \text{mot} \qquad H_1:\ \ \beta_2 > -0.1$$

Vi kan nå ikke bruke testobservatoren eller p-verdien gitt i datautskriften siden de er for en test hvor nullhypotesen er at $\beta_2 = 0$. Vi har imidlertid fra datautskriften all informasjon vi trenger for å selv utføre testen. Under nullhypotesen er (pensum/formelark):

$$T = \frac{\hat{\beta}_2 - \beta_{2_0}}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - (-0.1)}{SE(\hat{\beta}_2)} \sim t(n-k-1) = t(128-2-1) = t(125)$$

Vi har da at vi forkaster H_0 dersom $T > t_{0.05,125} = 1.657$. Observert:

$$t = \frac{-0.086146 + 0.1}{0.009114} = 1.52 < 1.657$$

Dvs, vi $\underline{\underline{\text{forkaster ikke } H_0}}$. Vi har altså ikke grunnlag for å konkludere at reduksjonen er saktere enn -0.1 per uke.

c) Residualet til en observasjon er definert som $\underline{\epsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}}$ og for den angitte observasjonen blir dette $\epsilon_i = y_i - 32.4 + 0.083 \cdot x_{1i} + 0.086 \cdot x_{2i} = 13.6 - 32.4 + 0.083 \cdot 180 + 0.086 \cdot 48 = \underline{\underline{0.26}}$.

Det første residualplottet er et plott av residualene mot x_1 (temperatur) og fra dette plottet kan vi sjekke om antagelsen om linær sammenheng mellom x_1 (temperatur) og forventningen til Y holder, og om antagelsen om lik varians i Y for alle x_1-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_1-verdier). Her ser det ut for å være litt mønster i residualplottet. Ved den laveste temperaturen ligger hovedtyngden av residualene under 0, og ved den neste høyeste temperaturen ligger nesten alle residualene over 0. Kanskje er det også litt forskjell i variasjonen ved ulike temperaturer (minst variasjon ved nest høyeste temperatur). Den andre residualplottet er et tilsvarende plott av residualene mot x_2 (tid). Dette plottet ser noe bedre ut, men kanskje litt stigende tendens mot slutten. Det tredje plottet er et plott av residualene mot estimert regresjonslinje (dvs plott av r_i mot $\hat{y}_i$). Dette plottet brukes primært til å sjekket om antagelsen om konstant varians ser ok ut. Konstant varians ser noenlunde ok ut, men vi ser også er en tendens til et ikke-lineært mønster med lave verdien for små og stor $\hat{y}$-verdier og høye verdier i midten. Det siste plottet er et normalplott som vi bruker til å sjekke om antagelsen om normalfordelt feilledd er ok. Punktene i plottet faller noenlunde på rett linje (med noen moderate unntak i endene) så antaglsen om normalfordeling ser noenlunde ok ut.

Det kan se ut for at der er noen ikke-lineære sammenhenger i dataene som den tilpassede modellen ikke har klart å fange opp. Mulig forbedringer å prøve kan derfor være f.eks. polynom eller transformasjoner av x_1 (temperatur), kanskje også polynom eller transformasjon av x_2 (tid), eller evt transformasjon av y (isoleringsevne).

Dersom rekkefølgen eksperimentene som gav dataene ble utført på er kjent kunne vi plotte residualene mot rekkefølgen for å se om der eventuelt er noen avhengigheter over tid.