Løsning eksamen 13. mai 2016

Oppgave 1:

a) Vi har en situasjon karakterisert ved:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" - flere studenter som er for omlegging eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, p - samme sannsynlighet p for å treffe en student som er for ved hver ny spørring.
- Enkeltforsøkene er uavhengige - uavhengige studenter pga de velges ut tilfeldig fra en stor gruppe.
- Et bestemt antall, n enkeltforsøk - et bestemt antall studenter som spørres.

Dermed har vi at X="antall studenter som er for omlegging" er binomisk fordelt med parametre n og p. Eventuelt kan man argumentere med at X strengt tatt er hypergeometrisk fordelt, men kan med god presisjon tilnærmes til binomisk fordeling så lenge n er mindre enn 10% av studentmassen.

Med $X \sim \text{Bin}(6, 0.4)$ får vi:

$$\begin{aligned} P(X > 3) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= \binom{6}{4}(0.4)^4(1-0.4)^{6-4} + \binom{6}{5}(0.4)^5(1-0.4)^{6-5} + \binom{6}{6}(0.4)^6(1-0.4)^{6-6} \\ &= 0.1382 + 0.0369 + 0.0041 = \underline{\underline{0.18}} \end{aligned}$$

Siden $np(1-p) = 60 \cdot 0.4 \cdot (1-0.4) = 14.4 > 5$ kan vi bruke tilnærming til normalfordeling:

$$\begin{aligned} P(X > 30) &= 1 - P(X \leq 30) \approx 1 - P(Z \leq \frac{30 + 0.5 - \text{E}(X)}{\sqrt{\text{Var}(X)}}) = 1 - P(Z \leq \frac{30 + 0.5 - np}{\sqrt{np(1-p)}}) \\ &= 1 - P(Z \leq \frac{30 + 0.5 - 60 \cdot 0.4}{\sqrt{60 \cdot 0.4 \cdot 0.6}}) = 1 - P(Z \leq 1.71) = 1 - 0.9564 = \underline{\underline{0.044}} \end{aligned}$$

(Om man utelater heltallskorreksjonen +0.5, får man enten svaret 0.057 dersom man starer ut som over, eller svaret 0.033 dersom man starter ut med $P(X > 30) = 1 - P(X < 31)$).

Det er logisk at sannsynligheten er lavest i den siste situasjonen pga vi spørr mange flere studenter. Jo flere studenter vi spør jo lavere er sannsynligheten for å få en observert andel som avviker betydelig fra reell andel. Når reell andel er 0.4 er det derfor mye lavere sannsynlighet for å få en observert andel på over 0.5 når man spørr 60 studenter enn når man spørr 6.

b) Estimat av p: $\hat{p} = 69/214 = \underline{\underline{0.322}}$. 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 $\hat{p} = 0.322$, $n = 214$ og $z_{\alpha/2} = z_{0.05} = 1.645$ gir dette tilnærmet 90% konfidensintervall for p:

$$[0.322 - 1.645\sqrt{\frac{0.322(1-0.322)}{214}}, \ \ 0.322 + 1.645\sqrt{\frac{0.322(1-0.322)}{214}}] = \underline{\underline{[0.27 \ , \ \ 0.37]}}$$

Tallene fra evalueringen i STA100 er neppe representative for hva alle studenter ved UiS mener. Det er slett ikke sikkert at de som tar STA100 mener det samme om dette som resten av studentmassen ved UiS. Videre er det heller ikke gitt at vi har data fra en representativ del av de som tar STA100. Det var frivillig å svare på undersøkelsen, mindre en halvparten av studentene i STA100 svarte og det kan hende at de som ikke svarte har et annet syn på spørsmålet enn de som svarte. Videre har vi også aspektet at de som svarte på evalueringen svarte på hva de mente om omlegging til videoforelesninger i STA100, og ikke om omlegging til videoforelesninger generelt. Dvs, vi kan ikke generalisere estimatet og konfidensintervallet til å si noe om hva hele studentmassen ved UiS mener om omlegging til videoforelesninger. Det gir kun en pekepinn på hva de aktive studentene i faget mener om spørsmålet i forhold til undervisningen i STA100.

Oppgave 2:

a) La M være antall pumper som feiler i løpet av en måned. Fra opplysningene i oppgaven vil da M være Poissonfordelt med $\lambda = 0.4$ og $t = 1$.

$$\begin{aligned} P(M=0) &= \frac{0.4^0}{0!}e^{-0.4} = 0.6703 = \underline{\underline{0.67}} \\ P(M<3) &= P(M=0)+P(M=1)+P(M=2) = 0.6703 + \frac{0.4^1}{1!}e^{-0.4} + \frac{0.4^2}{2!}e^{-0.4} \\ &= 0.6703 + 0.2681 + 0.0536 = \underline{\underline{0.992}} \end{aligned}$$

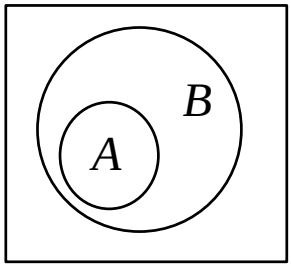
Evt kan man bruke tabell.

La Y være antall pumper som feiler i løpet av 12 måneder. Da vil Y være Poissonfordelt med $\lambda = 0.4$ og $t = 12$, dvs forventning $\lambda t = 0.4 \cdot 12 = 4.8$.

$$P(Y>6) = 1 - P(Y \leq 6) \overset{\text{tabell}}{=} 1 - 0.791 = \underline{\underline{0.21}}$$

Evt kan man summere sannsynlighetene for alle muligheter fra 0 til 6 i stedet for å bruke tabellen.

b) Merk at A er en delmengde av B slik at:



$$\begin{aligned} P(A) &= P(M \geq 3) = 1 - P(M<3) = 1 - 0.992 = \underline{\underline{0.008}} \\ P(B) &= P(M \geq 1) = 1 - P(M=0) = 1 - 0.67 = \underline{\underline{0.33}} \\ P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.008}{0.33} = \underline{\underline{0.024}} \end{aligned}$$

Hendelsene A og B er $\underline{\underline{\text{ikke uavhengige}}}$. Vi ser at $P(A|B) \neq P(A)$, og de kan heller ikke være uavhengige siden A er en delmengde av B.

c)

$$\begin{aligned} \mathrm{E}(X) &= \sum_x xP(X=x) = 0 \cdot 0.670 + 1 \cdot 0.268 + 2 \cdot 0.054 + 3 \cdot 0.008 = \underline{\underline{0.4}} \\ \mathrm{E}(X^2) &= \sum_x x^2 P(X=x) = 0^2 \cdot 0.670 + 1^2 \cdot 0.268 + 2^2 \cdot 0.054 + 3^2 \cdot 0.008 = 0.556 \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 0.556 - 0.4^2 = \underline{\underline{0.396}} \end{aligned}$$

Når $X < 3$ er den månedlige kostnaden $15X$ og når $X = 3$ er kostnaden $3 \cdot 15 + 55 = 100$. Dvs

$$\mathrm{E}(W) = \sum_x w(x)P(X=x) = 15 \cdot 0 \cdot 0.670 + 15 \cdot 1 \cdot 0.268 + 15 \cdot 2 \cdot 0.054 + 100 \cdot 0.008 = \underline{\underline{6.44}}$$

Oppgave 3:

a)

$$\begin{aligned} P(X > 30) &= 1 - P(\frac{X-27}{2} < \frac{30-27}{2}) = 1 - P(Z < 1.50) = 1 - 0.9332 = \underline{\underline{0.067}} \\ P(25 < X < 30) &= P(X < 30) - P(X < 25) \\ &= P(\frac{X-27}{2} < \frac{30-27}{2}) - P(\frac{X-27}{2} < \frac{25-27}{2}) \\ &= P(Z < 1.50) - P(Z < -1.00) = 0.9332 - 0.1587 = \underline{\underline{0.77}} \end{aligned}$$

La $Y = X_1 + X_2 + X_3 + X_4$ være summert funksjonstid. Da er $\mathrm{E}(Y) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \mathrm{E}(X_3) + \mathrm{E}(X_4) = 4 \cdot 27 = 108$ og $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \mathrm{Var}(X_3) + \mathrm{Var}(X_4) = 4 \cdot 2^2 = 16$ (pga uavhengighet). Siden Y er en lineærkombinasjon av normalfordelte variable vil Y også være normalfordelt og vi får da:

$$P(Y > 100) = 1 - P(\frac{Y-108}{\sqrt{16}} < \frac{100-108}{\sqrt{16}}) = 1 - P(Z < -2.00) = 1 - 0.0228 = \underline{\underline{0.977}}$$

b) Vi har uavhengige normalfordelte målinger med kjent standardavvik $\sigma = 2$ og skal teste:

$$H_0: \ \mu \leq 30 \qquad \text{mot} \qquad H_1: \ \mu > 30$$

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

$$Z = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{Var}(\overline{X})}} = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - 30}{2/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Med signifikansnivå $\alpha = 0.05$ forkaster vi H_0 dersom $Z \geq z_{0.05} = 1.645$. Observerte data gir:

$$z_{obs} = \frac{31.1 - 30}{2/\sqrt{35}} = 3.25 > 1.645.$$

Dvs. utfallet er i forkastningsområdet. Konklusjon: Vi $\underline{\underline{\text{forkaster } H_0}}$. Vi kan konkludere at forventet funksjonstid er over 30 timer.

$$p - \text{verdi} = P(Z > z_{obs}) = P(Z > 3.25) = 1 - P(Z < 3.25) = 1 - 0.9994 = \underline{\underline{0.0006}}$$

c) Vi har her en uparet sammenligning av to utvalg (to belastningstestsituasjoner). I tillegg til antagelsene om normalfordeling og uavhengighet gjort i oppgaveteksten antar vi at $\underline{\underline{\text{variansen er like store i begge situasjonene}}}$.

Et $(1-\alpha)100\%$ konfidensintervall for $\mu_X - \mu_Y$ er da gitt ved:

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2, n_X+n_Y-2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}},\ \bar{X} - \bar{Y} + t_{\alpha/2, n_X+n_Y-2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}\right]$$

Her er:

$$\begin{aligned} s_p^2 &= \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2} = \frac{(16-1)6.4^2 + (18-1)8.1^2}{16+18-2} = 54.055 \\ s_p &= \sqrt{54.055} = 7.35 \end{aligned}$$

Videre har vi $t_{\alpha/2, n_X+n_Y-2} = t_{0.025,32} = 2.037$ og vi får da følgende 95% konfidensintervall for $\mu_x - \mu_y$:

$$\left[187.3 - 196.0 - 2.037 \cdot 7.35\sqrt{\frac{1}{16} + \frac{1}{18}}, 187.3 - 196.0 + 2.037 \cdot 7.35\sqrt{\frac{1}{16} + \frac{1}{18}}\right] = \underline{\underline{[-13.8,\ -3.6]}}$$

Siden $\mu_X - \mu_Y = 0$ ikke er inneholdt i konfidensintervallet vil vi på 5% nivå forkaste den tosidig hypotesetesten av nullhypotesen $\mu_X - \mu_Y = 0$ mot alternativet $\mu_X - \mu_Y \neq 0$. Dvs vi $\underline{\underline{\text{forkaster } H_0}}$ og konkluderer at forventet funksjonstid i de to situasjonen er ulik.

Oppgave 4:

a) $\hat{y} = 55 + 1.43 \cdot x_1 + 0.63 \cdot x_2 + 0.15 \cdot x_3$. I følge denne ligningen øker forventet avgitt varme med 1.43 når x_i øker med 1.

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 0 \qquad \text{mot} \qquad H_1: \ \text{minst en } \beta_i \neq 0$$

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

$$F = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MSR}{MSE} \sim F(\alpha, k, n-k-1)$$

og vi forkaster nullhypotesen dersom F blir stor. Fra den oppgitte variansanalysetabellen ser vi at p-verdien for testen er $0.00000 < 0.05$, dvs vi forkaster H_0 på 5% nivå. Dette betyr at de tre kjemiske variablene har samlet sett innflytelse på avgitt varme.

Vi trenger x_3 i modellen dersom $\beta_3 \neq 0$, dvs vi skal teste:

$$H_0: \ \beta_3 = 0 \qquad \text{mot} \qquad H_1: \ \beta_3 \neq 0$$

Fra datautskriften leser vi at p-verdien for testen er $0.54 > 0.05$, dvs vi forkaster ikke H_0 på 5% nivå. Dette betyr at vi ikke trenger ha med x_3 i modellen. Når vi kjenner x_1 og x_2 har ikke x_3 påvirkning på mengden avgitt varme.

b) I den første modellen (med x_3) er $r^2 = 0.900$ og r^2-justert=0.882 og i den andre modellen (uten x_3) er $r^2 = 0.898$ og r^2-justert=0.886. Dvs forskjellene er marginale for både R^2 og R^2-justert, og da foretrekker vi vanligvis modellen med færrest variabler - dvs modellen uten x_3.

Merk at R^2 er litegrann høyere for modellen med flest variabler, R^2 øker alltid når vi tar med flere x-variabler. Når vi som her sammenligner modeller med ulike antall variabler bør vi derfor heller bruke R^2-justert som "straffer" det å ta med for mange variabler. Og vi ser at R^2-justert er litegrann lavere for modellen uten x_3. Konklusjonen er at modellen uten x_3 er best, ut fra lik/marginalt lavere R^2-justert.

$$H_0: \ \beta_1 \leq 1 \qquad \text{mot} \qquad H_1: \ \beta_1 > 1$$

Vi har fra penusm/formelarket at

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t(n-k-1)$$

Med $k = 2$ x-variabler og under $H_0: \ \beta_1 = 1$ får vi da at

$$T = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} \sim t(n-3)$$

og vi forkaster da H_0 dersom $T \geq t_{\alpha,n-3} = t_{0.05,17} = 1.74$.

Observert: $t_{obs} = \frac{1.359-1}{0.179} = 2.01$

Dvs, vi forkaster H_0 på 5% nivå, dataene gir grunn til å konkludere at forventet avgitt varme øker med mer enn 1 når x_1 øker med 1. Eller sagt på en annen måte, β_1, regresjonskoeffisienten for x_i, er større enn 1.