Løsning eksamen 14. mai 2020

Oppgave 1

a) Antall positive tester er binomisk fordelt dersom:

- Flere enkeltforsøk som hvert resulterer i "suksess" eller ikke "suksess" - flere prøver som er positive eller ikke.
- Sannsynligheten for "suksess" er den samme i alle enkeltforsøk, p - må være samme sannsynlighet p for hver prøve for å være positiv.
- Enkeltforsøkene er uavhengige - må være uavhengig fra prøve til prøve om den er positiv eller ikke.
- Et bestemt antall, n enkeltforsøk - et bestemt antall prøver som testes.

Dermed disse betingelsene er oppfylte er X="antall prøver som er positive" binomisk fordelt med parametre n og p.

Estimat av p : $\hat{p} = 59/1311 = \underline{\underline{0.045}}$.

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 $\hat{p} = 0.0045$, $n = 1311$ og $z_{\alpha/2} = z_{0.025} = 1.96$ gir dette tilnærmet 95% konfidensintervall for p :

$$[0.045 - 1.96\sqrt{\frac{0.045(1-0.045)}{1311}},\ \ 0.045 + 1.96\sqrt{\frac{0.045(1-0.045)}{1311}}] = \underline{\underline{[0.034\ ,\ \ 0.056]}}$$

b) Antall negative finner vi ved å ta det totale antallet og trekke fra antall positive. Vi får da tabellen:

	positive	negative
uke 1	59	1252
uke 2	40	1116

Første skritt i testen er å regne ut forventet antall i hver celle i tabellen under antagelsen om lik fordeling i gruppene. Vi trenger da også rad- og kolonnesummene. Disse og de forventede verdiene er gitt i tabellen under (de forventede verdiene står i parantesene i hver celle). For eksempel finner vi forventet verdi for kombinasjonen "uke 1' og "positiv" som: $1311 \cdot (99/2467) = 64.1$ og for kombinasjonen "uke 2" og "negativ" som: $1156 \cdot (2368/2467) = 107.1$.

	positive	negative	totalt
uke 1	59 (52.6)	1252 (1258.4)	1311
uke 2	40 (46.4)	1116 (1109.6)	1156
totalt	99	2368	2467

Testobservatoren blir da:

$$\begin{aligned}\mathcal{Q} &= \sum_{\text{alle celler}} \frac{(\text{observert-forventet})^2}{\text{forventet}} \\ &= \frac{(59-52.6)^2}{52.6} + \frac{(1252-1258.4)^2}{1258.4} + \frac{(40-46.4)^2}{46.4} + \frac{(1116-1109.6)^2}{1109.6} = 1.73\end{aligned}$$

Verdien på testobservatoren skal vi sammenligne med 5% kvantilen i kjikvadratfordelingen med parameter (frihetsgrader) $(r-1)\cdot(k-1) = (2-1)\cdot(2-1) = 1$. Fra kjikvadrat-tabellen finner vi at denne kvantilen har verdi 3.84. Siden $\mathcal{Q} = 1.73 < 3.84$ blir konklusjonen av vi ikke forkaster nullhypotesene om like andeler de to ukene. Dvs, vi kan ikke konkludere at der er noen generell forskjell i andel smittede i gruppen med symptomer mellom de to ukene.
Evt kan testen utføres i R ved å legge inn krysstabellen i R og bruke `chisq.test`-funksjonen.

c) Dersom antagelsens spesifisert i punkt a) holder er antall positive prøver blant de n prøvene binomisk fordelt. Vi har at binomisk fordeling kan tilnærmes med Poisson-fordeling med $\lambda = np$ (og $t = 1$) dersom $n > 10$ og $p < 0.10$, noe vi ser er oppfylt i de to neste spørsmålene i punktet.

Med $n = 200$ og $p = 0.03$ blir $\lambda = 200 \cdot 0.03 = 6$. Vi får da:

$$\begin{aligned} P(Y = 4) &= \frac{6^4}{4!}e^{-6} = \underline{\underline{0.13}} \\ P(Y \geq 4) &= 1 - P(Y \leq 3) \overset{\text{tabell}}{=} 1 - 0.151 = \underline{\underline{0.85}} \end{aligned}$$

Med $n = 1156$ og $p = 0.03$ blir Y Poisson-fordelt med $\lambda = 1156 \cdot 0.03 = 34.68$. Siden $\lambda t = 34.68 \cdot 1 > 10$ kan vi bruke tilnærming til normalfordeling. Husk at i Poissonfordeling er $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda t$.

$$P(X \geq 40) = 1 - P(X \leq 39) = 1 - P(Z \leq \frac{39 + 0.5 - 34.68}{\sqrt{34.68}}) = 1 - P(Z \leq 0.82) = 1 - 0.7939 = \underline{\underline{0.21}}$$

(Det er ikke farlig om heltallskorreksjonen, +0.5, droppes. Svaret blir da enten 0.23 eller 0.18 alt etter om man tar $P(X \geq 40) = 1 - P(X \leq 39)$ eller $P(X \geq 40) = 1 - P(X < 40)$ som utgangspunkt.)

Kan alternativt (og litt mer presist) bruke tilnærming direkte fra binomisk til normalfordeling. Oppgavene i dette punktet kan også løses i R ved å bruke `dbinom` og `pbinom` (gir eksakte svar) eller `dpois` og `ppois` (gir tilnærmede svar).

d) Vi lar X betegne antall smittede blant n som testes. Av samme grunner som innledningsvis i punkt a) vil da X være binomisk fordelt,og det oppgis i oppgaven at vi skal anta at tilnærming til normalfordeling er OK (som er tilfelle når $np(1-p) > 5$. Vi har da fra pensum at $\hat{p} = X/n$ også vil være tilnærmet normalfordelt med $\mathrm{E}(\hat{p}) = p$ og $\mathrm{E}(\hat{p}) = p(1-p)/n$. Dermed:

$$\begin{aligned} P(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) &= P(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}) \approx P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= P(Z < z_{\alpha/2}) + P(Z < -z_{\alpha/2}) = 1 - \alpha/2 + \alpha/2 = 1 - \alpha \end{aligned}$$

Sannsynligheten for at en andel $\hat{p}$ faller *utenfor* intervallet $[p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}]$ blir da $\underline{\alpha}$.

For Z har vi:

- Gjentatte delforsøk som gir "suksess"/ikke "suksess" - gjentatte tester hvor andelen enten faller utenfor intervallet eller ikke.
- Lik sannsynlighet i alle delforsøk - samme sannsynlighet α for at resultatet av vil falle utenfor intervallet for hver gang.
- Uavhengige delforsøk - uavhengig resultat fra gang til gang pga det er tilfeldig valgte personer.
- Gjentar delforsøkene inntil første suksess - tester til man første gang får et resultat utenfor grensene.

Dvs, alle betingelsene for geometrisk fordeling er oppfylte og vi har dermed at Z er geometrisk fordelt med parameter $p = \alpha$.

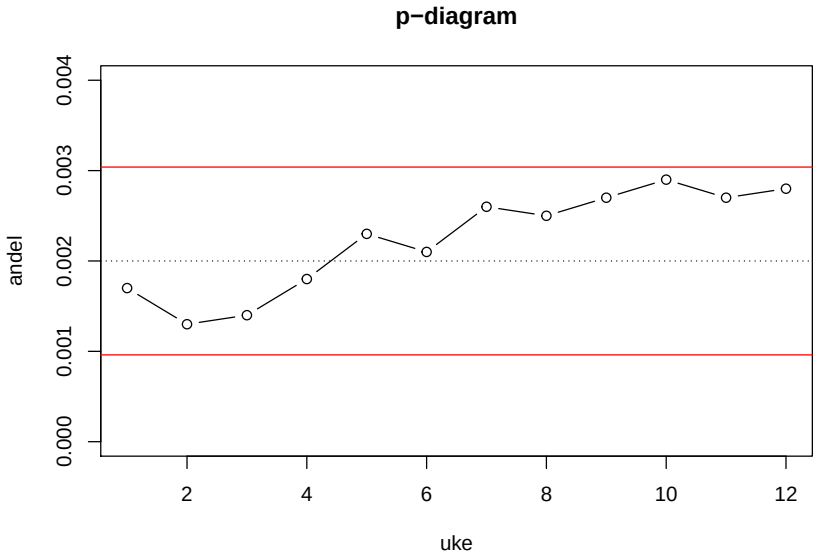
I geometrisk fordeling har vi at $\mathrm{E}(Z) = 1/p$. ARL er forventet tid mellom hver gang en stikkprøve faller utenfor grensene når prosessen er i kontroll (her når sann andel virkelig er p), og siden sannsynligheten for en verdi utenfor grensene er $p = \alpha$ blir $\mathrm{ARL} = \mathrm{E}(Y) = 1/p = 1/\alpha$.

e) Med $ARL = 50$ blir $\alpha = 1/50 = 0.02$, og dermed $z_{\alpha/2} = z_{0.01} = 2.326$. Merk at siden vi i denne situasjonen kjenner p trenger vi ikke estimere p fra dataene. Senterlinja i diagrammet blir da bare $p = 0.002$ og for kontrollgrensene får vi:

Øvre kontrollgrense: $p + 2.326\sqrt{p(1-p)/n} = 0.002 + 2.326\sqrt{0.002(1-0.002)/10000} = \underline{\underline{0.0030}}$.

Nedre kontrollgrense: $p - 2.326\sqrt{p(1-p)/n} = 0.002 + 2.326\sqrt{0.002(1-0.002)/10000} = \underline{\underline{0.0010}}$.

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



Vi ser at ingen av stikkprøvene går utenfor kontrollgrensene. Men samtidig ser vi en jevnt økende tendens over perioden, og alle de åtte siste registreringene er over senterlinjen. Dette er en indikasjon på en økende tendens. I praksis er en svakhet med $\hat{p}$-diagram (og $\bar{x}$- og s-diagram) at det ikke er godt egnet til å fange opp slike mindre men vedvarende endringer og i praksis vil man derfor trolig velge en annen type diagram i en situasjon som dette (f.eks. CUSUM eller EWMA som ikke er med i pensum i STA100).

Oppgave 2

a) Vi lar X være rekkevidden, og vi får da:

$$\begin{aligned} P(X > 234) &= 1 - P(\frac{X-245}{25} < \frac{234-245}{25}) = 1 - P(Z < -0.44) = 1 - 0.3300 = \underline{0.67} \\ P(X < 252) &= P(\frac{X-245}{25} < \frac{252-245}{25}) = P(Z < 0.28) = \underline{\underline{0.61}} \end{aligned}$$

Evt går det her an å bruke `pnorm`-funksjonen i R.

Fra det vi har regnet ut så lang vet vi at sannsynligheten for å ha for kort rekkevidde er 1-067=0.33 når strekningen er 234 km og 0.61 når det er 252 km. Det er totalt 6 muligheter for hvilken tur hun ikke har nok rekkevidde, fire muligheter for turene med omkjøring og to muligheter for turene uten omkjøring. Sannsyunligheten blir da:

$$4 \cdot 0.33 \cdot 0.67^3 \cdot 0.39^2 + 0.67^4 \cdot 2 \cdot 0.61 \cdot 0.39 = \underline{\underline{0.16}}$$

b) La μ betegne forventet rekkevidde. Vi skal da teste

$$H_0: \ \mu \geq 310 \qquad \text{mot} \qquad H_1: \ \mu < 310$$

Vi antar at rekkevidden er normalfordelt (med ukjent μ og σ), og at de ulike målingene av rekkevidden som er gjort er uavhengige. Dersom H_0 er korrekt er da

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

Med signifikansnivå 5%, dvs $\alpha = 0.05$, og $n = 10$ forkaster vi H_0 dersom $T \leq -t_{0.05,9} = -1.833$.
Fra oppgitt info får vi $\bar{x} = 3042/10 = 304.2$ og $s = \sqrt{\sum_{i=1}^{10}(x_i - \bar{x})^2/(n-1)} = \sqrt{1005.6/9} = 10.57$
Observert: $t = \frac{304.2-310}{10.57/\sqrt{10}} = -1.74$
Siden $-1.74 > -1.833$ blir konklusjonen at vi ikke forkaster H_0. Dataene gir ikke grunnlag for å konkludere, på 5% nivå, at forventet rekkevidde er mindre enn 310.

c) Her kan vi tolke α som forventet rekkevidde ved 0 grader og β som endring i forventet rekkevidde når temperaturen øker en grad.

Estimert regresjonsline: $\hat{y} = \hat{\alpha} + \hat{\beta}x = 245 + 2.304 \cdot x$.
Ved $x = 15$ grader får vi $\hat{y} = 245 + 2.304 \cdot 15 = 280$.

Med testen $H_0: \ \beta = 0$ mot $H_1: \ \beta \neq 0$ avgjør vi om det er sammenheng mellom x- og Y-variabelen, dvs mellom temperatur og rekkevidde. Vi ser fra dataustksriften at p-verdien$=4.8 \cdot 10^{-10} < 0.05$ dvs vi forkaster H_0 og konkluderer med at der er sammenheng mellom temperatur og rekkevidde.

Vi ser fra plottet at korrelasjonen er positiv (økende y-verdi for økende x-verdi, ser også at estimert β-verdi er positiv) og vi ser fra dataustksriften at $R^2 = 0.780$, korrelasjonen er dermed $\sqrt{0.780} = 0.88$.

Estimert forventet forskjell blir: $8\hat{\beta} = 8 \cdot 2.3036 = 18.4$.Vi ser videre fra datautskriften at et 95% konfidensintervall for β er $[1.811, \ 2.796]$. Et 95% konfidensintervall for 8β er da $[8{\cdot}1.811, \ 8{\cdot}2.796] = [14.5, \ 22.4]$.

Oppgave 3

a)

$$\begin{aligned}
\mathrm{E}(X) &= \sum_x xP(X=x) = 1 \cdot 0.04 + 2 \cdot 0.08 + 3 \cdot 0.21 + 4 \cdot 0.35 + 5 \cdot 0.32 = 3.83 \\
\mathrm{E}(X^2) &= \sum_x x^2P(X=x) = 1^2 \cdot 0.04 + 2^2 \cdot 0.08 + 3^2 \cdot 0.21 + 4^2 \cdot 0.35 + 5^2 \cdot 0.32 = 15.85 \\
\Rightarrow \mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = 15.85 - 3.83^2 = 1.18 \\
\mathrm{E}(Y) &= 1 \cdot 0.03 + 2 \cdot 0.07 + 3 \cdot 0.21 + 4 \cdot 0.36 + 5 \cdot 0.33 = 3.89
\end{aligned}$$

Vi finner $P(Y > X)$ ved å summere sannsynligheten for alle mulige kombinasjoner av X og Y der $Y > X$ (siden vi trekke tilfeldige studenter er X og Y uavhengige):

$$\begin{aligned}
P(Y > X) &= P(X = 1 \cap Y > 1) + P(X = 2 \cap Y > 2) + P(X = 3 \cap Y > 3) + P(X = 4 \cap Y > 4) \\
&\overset{\text{uavh.}}{=} P(X=1)P(Y>1) + P(X=2)P(Y>2) + P(X=3)P(Y>3) + P(X=4)P(Y=5) \\
&= 0.04 \cdot (1 - 0.03) + 0.08 \cdot (1 - 0.03 - 0.07) + 0.21 \cdot (0.36 + 0.33) + 0.35 \cdot 0.33 \\
&= 0.37
\end{aligned}$$

(Videre vil $P(Y = X) = 0.04 \cdot 0.03 + 0.08 \cdot 0.07 + 0.21 \cdot 0.21 + 0.35 \cdot 0.36 + 0.32 \cdot 0.33 = 0.28$ og dermed $P(Y < X) = 1 - 0.37 - 0.28 = 0.35$.)

b) Siden vi skal avgjøre om om μ_X og μ_Y er forskjellige er det naturlig å formulere problemstillingen som følgende hypotesetest:

$$H_0: \ \mu_X = \mu_Y \qquad \text{mot} \qquad H_1: \ \mu_X \neq \mu_Y$$

Dersom vi antar at svarene fra de ulike studentene er uavhengige fra en fordeling med forventningsverdi μ_X og standardavvik σ_X ved UiS og uavhengige fra en fordeling med forventningsverdi μ_Y og standardavvik σ_Y ved det andre lærestedet følger det fra sentralgrenseteoremet at $\bar{X} \approx \mathrm{N}(\mu_X, \sigma_X/\sqrt{n_X})$ og $\bar{Y} \approx \mathrm{N}(\mu_Y, \sigma_Y/\sqrt{n_Y})$. Vi får da videre at:

$$Z = \frac{\bar{X} - \bar{Y} - \mathrm{E}(\bar{X} - \bar{Y})}{\sqrt{\mathrm{Var}(\bar{X} - \bar{Y})}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y})}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} \approx \mathrm{N}(0,1)$$

Under H_0 er $\mu_X = \mu_Y$ og tilnærmingen til normalfordeling gjelder fremdeles når σ_X og σ_Y byttes ut med estimatorene $\hat{\sigma}_X$ og $\hat{\sigma}_Y$ (merk at vi her ikke får t-fordeling pga dataene er ikke normalfordelte). Vi ender dermed opp med testobservatoren:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_X^2/n_X + \hat{\sigma}_Y^2/n_Y}}$$

Siden ikke er oppgitt noe nivå i oppgaven må vi selv bestemme det, og det er da rimelig å velge det vanlige nivået $\alpha = 0.05$. Vi får da at vi forkaster H_0 dersom $Z < -z_{\alpha/2} = -z_{0.025} = -1.96$ eller $Z > 1.96$.

Observert:

$$z_{\text{obs.}} = \frac{3.83 - 3.89}{\sqrt{1.09^2/1384 + 1.04^2/987}} = -1.36$$

Konklusjonen blir dermed at vi ikke forkaster H_0, dataene gir ikke grunnlag for å konkludere at det er forskjell i forventet fornøydhet mellom de to lærestedene.

Antagelsene vi har gjort er at studentene som svarer er uavhengige fra en fordeling med forventningsverdi μ_X og standardavvik σ_X ved UiS og uavhengige fra en fordeling med forventningsverdi μ_Y og standardavvik σ_Y ved det andre lærestedet. Antagelsen om uavhengighet mellom studentene er antagelig noenlunde rimelig, men det er uvisst om de studentene som har svart er representative for alle studenter. Vi har dermed ikke nødvendigvis et tilfeldig utvalg fra alle studenter, men fra en subgruppe som kan tenkes å ha en annen mening om påstanden vi har sett på her enn resten av studentmassen (og dermed f.eks. en annen μ_X og μ_Y en hele studentmassen de to lærestedene).

Vi ser at for UiS og det andre lærestedet vi har undersøkt, som har hhv et gjennomsnitt på 3.8 og 3.9, så er ikke forskjellen signifikant. Dvs, vi kan ikke konkludere at der er en forskjell i det hele tatt. Vi ser også fra punkt a) at forskjellen mellom en fordeling som gir forventningsverdi 3.8 og en fordeling som gir forventningsverdi 3.9 er helt marginal. Det samme vil gjelde mellom en fordeling som gir forventningsverdi 3.8 og en som gir forventningsverdi 4.0. Dvs, realiteten er at studentene ved UiS er omtrent like fornøyde som de fleste andre steder i landet, og de er godt fornøyde siden gjennomsnittskåren er 3.8 på en skala fra 1-5.