Løsningsforslag Eksamen i SIF 5505 Statistikk 9/1 1999

Oppgave 1

a)

$P(Y \leq 12) = P(\frac{Y-8}{2} \leq \frac{12-8}{2}) = \Phi(\frac{12-8}{2}) = \Phi(2) = \underline{\underline{0.977}}$

$P(Y > 5) = 1 - P(\frac{Y-8}{2} \leq \frac{5-8}{2}) = 1 - \Phi(-\frac{3}{2}) = 1 - (1 - \Phi(\frac{3}{2})) = \Phi(\frac{3}{2}) = \underline{\underline{0.933}}$

$P(5 < Y \leq 12) = P(Y \leq 12) - P(Y \leq 5) = 0.977 - (1 - 0.933) = \underline{\underline{0.910}}$

P.g.a. uavhengighet er:

$$\begin{aligned} P(\text{konsentrasjon for 8 kurer i } (5,12]) &= P(\text{konsentrasjon for en kur i } (5,12])^8 \\ &= P(5 < Y \leq 12)^8 \\ &= 0.91^8 = \underline{\underline{0.47}} \end{aligned}$$

b)

A_1 og A_2 er ikke disjunkte siden den ene hendelsen er delvis inneholdt i den andre ($5 < Y < 12$ er felles for begge hendelsene) dvs $A_1 \cap A_2 \neq \emptyset$.

A_1 og A_2 er ikke uavhengige. Dersom vi f.eks. vet at $Y > 5$ reduserer dette sannsynligheten for at $Y \leq 12$, dvs ikke uavhengige hendelser.

Eventuelt kan det vises ved regning at A_1 og A_2 ikke er uavhengige: Fra a) har vi at $P(A_1) \cdot P(A_2) = 0.977 \cdot 0.933 = 0.912$ og at $P(A_1 \cap A_2) = 0.910$. Dvs $P(A_1) \cdot P(A_2) \neq P(A_1 \cap A_2)$, dermed er A_1 og A_2 avhengige.

Eller enda enklere: $P(A_1 | A_2^C) = 1 \neq P(A_1)$. Dvs A_1 og A_2^C er avhengige og dermed er også A_1 og A_2 avhengige.

$\underline{\underline{A_3 = A_1 \cap A_2}}$

c)

Estimator: $\hat{\mu} = \underline{\underline{\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i}}$

Estimat: $\hat{\mu} = \frac{1}{8} \sum_{i=1}^{8} y_i = \frac{1}{8} \cdot 69.8 = \underline{\underline{8.725}}$

Konfidensinterval: Merk at i følge oppgaveteksten skal konfidensintervallet utledes - ikke bare settes opp.

Siden $\hat{\mu}$ er en lineærkombinasjon av uavhengige normalfordelte stokastiske variable vil $\hat{\mu}$ selv være normalfordelt.

$\mathrm{E}(\hat{\mu}) = \mathrm{E}(\frac{1}{n} \sum_{i=1}^{n} Y_i) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(Y_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$.

$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(\frac{1}{n} \sum_{i=1}^{n} Y_i) \overset{\text{(uavh.)}}{=} \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$.

Dvs vi kan ta utgangspunkt i at $\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$.

$$\begin{aligned} P(-u_{\alpha/2} \leq \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \leq u_{\alpha/2}) &= 1-\alpha \\ P(\hat{\mu} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}) &= 1-\alpha \end{aligned}$$

95% konf. int.:

$$\underline{\underline{[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}]}}$$

Med oppgitte data:

$$[8.725 - 1.96\frac{2}{\sqrt{8}}, 8.725 + 1.96\frac{2}{\sqrt{8}}] = \underline{\underline{[7.34, 10.11]}}$$

d)

Det er rimelig å ikke ha med noe konstantledd i regresjonsmodellen fordi dersom medisindosen velges lik null vil vi ikke få noen medisinkonsentrasjon ($x = 0 \Rightarrow Y = 0$).

For å sette opp rimelighetsfunksjonen (likelihoodfunksjonen) tar vi utgangspunkt i at residualene $E_1, \ldots, E_n$ er uavhengige og $\mathrm{N}(0, \sigma_E^2)$:

$$\begin{aligned} L(\beta; e_1, \ldots, e_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_E} e^{-\frac{1}{2}(\frac{e_i}{\sigma_E})^2} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_E} e^{-\frac{1}{2}(\frac{y_i-\beta x_i}{\sigma_E})^2} = \frac{1}{(2\pi)^{n/2}\sigma_E^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{y_i-\beta x_i}{\sigma_E})^2} \\ l(\beta) = \ln L(\beta) &= -\frac{n}{2}\ln(2\pi) - n\ln\sigma_E - \frac{1}{2}\sum_{i=1}^{n}(\frac{y_i-\beta x_i}{\sigma_E})^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} = -\sum_{i=1}^{n} \frac{y_i - \beta x_i}{\sigma_E^2}(-x_i) &= 0 \\ \sum_{i=1}^{n}(y_i x_i - \beta x_i^2) &= 0 \\ \beta &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \\ \hat{\beta} &= \underline{\underline{\frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}}} \end{aligned}$$

$\mathrm{E}(\hat{\beta}) = \mathrm{E}(\frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}) = \frac{\sum_{i=1}^{n} \mathrm{E}(Y_i) x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} \beta x_i x_i}{\sum_{i=1}^{n} x_i^2} = \underline{\underline{\beta}}$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right) &= \left(\frac{1}{\sum_{i=1}^n x_i^2}\right)^2 \mathrm{Var}\left(\sum_{i=1}^n Y_i x_i\right) \\
&\overset{\text{(uavh.)}}{=} \left(\frac{1}{\sum_{i=1}^n x_i^2}\right)^2 \sum_{i=1}^n x_i^2 \,\mathrm{Var}(Y_i) \\
&= \left(\frac{1}{\sum_{i=1}^n x_i^2}\right)^2 \sum_{i=1}^n x_i^2 \sigma_E^2 \\
&= \underline{\underline{\frac{\sigma_E^2}{\sum_{i=1}^n x_i^2}}}
\end{aligned}
$$

e) Tolkningen av et 95% prediksjonsintervall er at det er 95% sannsynlighet for at en ny observasjon Y_0, med kjent $x = x_0$, vil falle innenfor prediksjonsintervallet.

Predikert verdi for den nye observasjonen Y_0 er $\hat{\beta}x_0$. Vi tar utgangspunkt i differansen $\hat{\beta}x_0 - Y_0$ for å lage prediksjonsintervallet. Vi har at $\hat{\beta}x_0$ og Y_0 er uavhengige (siden Y_0 er en ny uavhengig observasjon), og:

$\mathrm{E}(\hat{\beta}x_0 - Y_0) = \mathrm{E}(\hat{\beta})x_0 - \mathrm{E}(Y_0) = \beta x_0 - \beta x_0 = 0$

$\mathrm{Var}(\hat{\beta}x_0 - Y_0) = \mathrm{Var}(\hat{\beta}x_0) + \mathrm{Var}(Y_0) = \frac{\sigma_E^2 x_0^2}{\sum_{i=1}^n x_i^2} + \sigma_E^2$

$\hat{\beta}x_0 - Y_0$ er en lineærkombinasjon av uavhengige normalfordelte variable og er følgelig normalfordelt.

$$
\begin{aligned}
P\left(-u_{\alpha/2} \le \frac{\hat{\beta}x_0 - Y_0}{\sqrt{\frac{\sigma_E^2 x_0^2}{\sum_{i=1}^n x_i^2} + \sigma_E^2}} \le u_{\alpha/2}\right) &= 1-\alpha \\
P\left(\hat{\beta}x_0 - u_{\alpha/2}\sqrt{\sigma_E^2 + \frac{\sigma_E^2 x_0^2}{\sum_{i=1}^n x_i^2}} \le Y_0 \le \hat{\beta}x_0 + u_{\alpha/2}\sqrt{\sigma_E^2 + \frac{\sigma_E^2 x_0^2}{\sum_{i=1}^n x_i^2}}\right) &= 1-\alpha
\end{aligned}
$$

Prediksjonsintervall: $[\hat{\beta}x_0 - u_{\alpha/2}\sigma_E\sqrt{1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}}, \hat{\beta}x_0 + u_{\alpha/2}\sigma_E\sqrt{1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}}]$

$\hat{\beta} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} = \frac{536.4}{436} = 1.23$

95% prediksjonsintervall for $x_0 = 8$:

$$
[1.23 \cdot 8 - 1.96 \cdot 2\sqrt{1 + \frac{8^2}{436}}, 1.23 \cdot 8 + 1.96 \cdot 2\sqrt{1 + \frac{8^2}{436}}] = \underline{\underline{[5.64, 14.04]}}
$$

Oppgave 2

a)

X vil være binomisk fordelt med $n = 100$ og $p = 0.01$ dersom følgende forutsetninger er oppfylte:

- Vi har n uavhengige forsøk/frimerke (dvs vi kan f.eks. ikke ha par av frimerke som henger sammen).
- For hvert frimerke registrerer vi om det er av varianten "ØRF" eller ikke.
- $P(\text{"ØRF"}) = 0.01$ for alle frimerke.

$P(X = 0) = \binom{100}{0} 0.01^0 0.99^{100} = 0.99^{100} = \underline{\underline{0.366}}$

$P(X = 1) = \binom{100}{1} 0.01^1 0.99^{99} = 0.370$

$P(X > 0) = 1 - P(X = 0) = 1 - 0.366 = 0.634$

$P(X > 1) = 1 - P(X = 0) - P(X = 1) = 1 - 0.366 - 0.370 = 0.264$

$P(X > 1|X > 0) = \frac{P(X>1 \cap X>0)}{P(X>0)} = \frac{P(X>1)}{P(X>0)} = \frac{0.264}{0.634} = \underline{\underline{0.416}}$

b)

For å kunne bruke normalfordelingstilnærmingen må vi ha $np > 5$ og $n(1 - p) > 5$. Her er $np = 4$, dvs normalfordelingstilnærmingen kan ikke brukes. (Dessuten er p nær 0, noe som også bidrar til å gjøre normaltilnærmingen dårlig.)

Hypotesetest: $H_0: \quad p = 0.01 \quad$ mot $H_1: \quad p < 0.01$

Vi bruker X som testobservator og vi forkaster H_0 dersom $X \leq k$, der k velges slik at $P(X \leq k|H_0) \leq \alpha = 0.05$. Ved samme antagelser som i a) er X binomisk fordelt med $n = 400$ og $p = 0.01$ under nullhypotesen.

$P(X \leq 0|p = 0.01) = P(X = 0|p = 0.01) = \binom{400}{0} 0.01^0 0.99^{400} = 0.99^{400} = 0.018$

$P(X \leq 1|p = 0.01) = P(X = 0|p = 0.01) + P(X = 1|p = 0.01) = 0.018 + \binom{400}{1} 0.01^1 0.99^{399} =$
0.090

Dvs siden $\alpha = 0.05$ velger vi $k = 0$. H_0 forkastes dersom $X \leq 0$.

Vi har observert $x = 0$, dvs $\underline{\underline{H_0 \text{ forkastes}}}$.

Alternativ løsning: p-verdi $= P(X \leq 0|p = 0.01) = 0.99^{400} = 0.018$. Siden p-verdien er mindre enn signifikansnivået på 5% $\underline{\underline{\text{forkastes } H_0}}$.

c)

Definerer først sannsynligheten q:

$$\begin{aligned} q = P(\text{Minst en "ØRF" i en 100-pakning}) &= 1 - P(\text{Ingen "ØRF" i en 100-pakning}) \\ &= 1 - (1-p)^{100} \end{aligned}$$

Med $p = 0.01$ blir $q = 0.634$.

For $Z < k$ er punktsannsynligheten til Z gitt ved:

$P(Z = 1) = q$

$P(Z = 2) = (1-q)q$

$P(Z = z) = (1-q)^{z-1}q, \quad 1 \leq z < k$

Dvs som en geometrisk fordeling, mens for $Z = k$ kan vi bruke følgende resonnement:

$P(Z = k) = P(\text{Ingen "ØRF" i de } k-1 \text{ første pakkene.}) = (1-q)^{k-1}$

Alternativt resonnement:

$$\begin{aligned} P(Z = k) = 1 - P(Z < k) = 1 - \sum_{z=1}^{k-1} P(Z = z) &= 1 - \sum_{z=1}^{k-1} (1-q)^{z-1} q \\ &= 1 - \sum_{i=0}^{k-2} (1-q)^{i} q \\ &\overset{\text{(geometrisk rekke)}}{=} 1 - q\frac{1-(1-q)^{k-1}}{1-(1-q)} \\ &= (1-q)^{k-1} \end{aligned}$$

Dvs punktsannsynligheten til Z er:

$$P(Z = z) = \begin{cases} (1-q)^{z-1}q & \text{for } 1 \leq z < k \\ (1-q)^{k-1} & \text{for } z = k \\ 0 & \text{ellers,} \end{cases}$$

der $q = 1 - (1-p)^{100}$

$\mathrm{E}(Z) = \sum_{z=1}^{4} zP(Z = z) = 1 \cdot q + 2 \cdot (1-q)q + 3 \cdot (1-q)^2 q + 4 \cdot (1-q)^3 \overset{(q=0.634)}{=} \underline{\underline{1.55}}$

Oppgave 3

Regner først ut den kumulative fordelingsfunksjonen til X:

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \quad \text{for} \quad x > 0$$

Skal finne sannsynlighetstettheten til $V = \max(X_1, X_2)$ og regner først ut fordelingsfunksjonen:

$$\begin{aligned} F_V(v) = P(\max(X_1, X_2) \leq v) &= P(X_1 \leq v \cap X_2 \leq v) \\ &\overset{uavh}{=} P(X_1 \leq v) P(X_2 \leq v) \\ &= F_X(v)^2 = (1 - e^{-\lambda v})^2 = 1 - 2e^{-\lambda v} + e^{-2\lambda v} \end{aligned}$$

Dvs. sannsynlighetstettheten til V blir:

$$f_V(v) = F'(v) = \underline{\underline{2\lambda e^{-\lambda v} - 2\lambda e^{-2\lambda v}}}$$

$$\begin{aligned} \mathrm{E}(V) = \int_0^\infty v(2\lambda e^{-\lambda v} - 2\lambda e^{-2\lambda v}) dv &= 2 \int_0^\infty v \lambda e^{-\lambda v} dv - \int_0^\infty v 2\lambda e^{-2\lambda v} dv \\ &= 2\frac{1}{\lambda} - \frac{1}{2\lambda} = \underline{\underline{\frac{3}{2\lambda}}} \end{aligned}$$

Vi har at $\mathrm{E}(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$, dvs. vi har at $\underline{\underline{\mathrm{E}(X) < \mathrm{E}(V) < 2\,\mathrm{E}(X)}}$ som ventet da V er den største av to X-er. Siden $V = \max(X_1, X_2)$ vil vi forvente at $\mathrm{E}(V) > \mathrm{E}(X)$ og at $\mathrm{E}(V) < \mathrm{E}(X_1 + X_2) = 2\,\mathrm{E}(X)$.

Oppgave 4

Eva ønsker å teste hypotesen:

$$H_0: \ \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0) \qquad \text{mot} \qquad H_1: \ \mu_1 < \mu_2 \quad (\mu_1 - \mu_2 < 0)$$

Siden variansen i kvinner og menns lønn er antatt lik tar vi utgangspunkt i testobservatoren:

$$T = \frac{\bar{X} - \bar{Y}}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

der $S^2_{pooled} = \frac{1}{n+n-2}((n-1)S^2_X + (n-1)S^2_Y) = \frac{1}{2n-2}(\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2)$. Vi har fra pensum at denne testobservatoren under H_0 (at $\mu_1 - \mu_2 = 0$) er t_{n+n-2}-fordelt. H_0 forkastes dersom $T < -t_{\alpha,2n-2}$ der

$$P(\text{forkaste } H_0 | H_0) = P(T < -t_{\alpha,2n-2}) = \alpha = 0.05$$

Vi har observert $s^2_{pooled} = \frac{1}{16-2}(39571.875 + 50550) = 6437.277 = 80.23^2$ som gir

$$t_{obs} = \frac{330.625 - 357.500}{80.23\sqrt{\frac{2}{8}}} = -0.670 > -t_{0.05,14} = -1.76$$

Dvs vi forkaster ikke H_0 på 5% nivå. Dataene gir ikke grunnlag for å påstå at firmaet driver med kvinnediskriminerende avlønning.