TMA4245 Statistikk
Eksamen desember 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

To hendelser: A=komponenten har en feil av type A, og B=komponenten har en feil av type B, og $P(B) = 0.09$, $P(A \mid B) = 0.5$ og $P(A \mid B') = 0.01$.

a)

$$\begin{aligned} P(A \cap B) &= P(A \mid B) \cdot P(B) = 0.5 \cdot 0.09 = 0.045 \\ P(A) &= P(A \cap B) + P(A \cap B') = P(A \cap B) + P(A \mid B') \cdot P(B') \\ &= P(A \cap B) + P(A \mid B') \cdot (1 - P(B)) = 0.045 + 0.01 \cdot (1 - 0.09) \\ &= 0.045 + 0.0091 = 0.054 \\ P(B \mid A) &= \frac{P(A \cap B)}{P(A)} = \frac{0.045}{0.054} = 0.83 \end{aligned}$$

Over: først definisjon av betinget sannsynlighet, så totalt sannsynlighet og til slutt Bayes' regel.

b) X er binomisk fordelt. Vi har en Bernoulliprosess (binomisk forsøksrekke)

- Vi velger komponenter fra produksjonen uavhengig av hverandre.
- For hver komponent undersøker vi om den er feilfri (suksess) eller ikke (fiasko).
- Sannsynligheten for suksess er 0.9 for alle komponenter.

og vi har gitt at $n = 20$ komponenter testes og vi lar X være antallet suksesser (feilfrie), som dermed vil være binomisk fordelt med $n = 20$ forsøk og suksess-sannsynlighet $p = 0.9$.

Sannsynligheter:

$$\begin{aligned} P(X = 19) &= \binom{20}{19} 0.9^{19}(1 - 0.9)^{20-19} = 0.270 \\ P(X > 15) &= 1 - P(X \leq 15) = 1 - 0.043 = 0.957 \end{aligned}$$

der $P(X \leq 15)$ er funnet på side 17 i Tabeller og formler i statistikk.

c) 90% konfidensintervall for p basert på at $Z = \frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}}$ er tilnærmet standard normalfordelt.

$$
\begin{aligned}
P(-z_{0.05} < Z < z_{0.05}) &= 0.90 \\
P(-z_{0.05} < \frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{0.05}) &= 0.90 \\
P(-z_{0.05}\sqrt{n\hat{p}(1-\hat{p})} < X - np < z_{0.05}\sqrt{n\hat{p}(1-\hat{p})}) &= 0.90 \\
P(-X - z_{0.05}\sqrt{n\hat{p}(1-\hat{p})} < -np < -X + z_{0.05}\sqrt{n\hat{p}(1-\hat{p})}) &= 0.90 \\
P(\frac{X}{n} - z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \frac{X}{n} + z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) &= 0.90 \\
P(\hat{p} - z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) &= 0.90 \\
P(\hat{p}_L < p < \hat{p}_U) &= 0.90
\end{aligned}
$$

Innsatt numerisk verdi for observert $\hat{p}$ og n kalles intervallet $[\hat{p}_L, \hat{p}_U]$ et 90% konfidensintervall for p.

Innsatt $n = 500$ og $x = 470$ feilfrie komponenter gir $\hat{p} = 0.94$.

$$
\begin{aligned}
\hat{p}_L &= 0.94 - 1.645 \cdot \sqrt{\frac{0.94 \cdot (1-0.94)}{500}} = 0.923 \\
\hat{p}_L &= 0.94 + 1.645 \cdot \sqrt{\frac{0.94 \cdot (1-0.94)}{500}} = 0.957
\end{aligned}
$$

Dette intervallet er et intervall der vi har 90% tillit til at vi finner den ukjente andelen feilfrie komponenter i produksjonen. Hvis vi gjentar forsøket med $n = 500$ komponenter mange ganger, så vil i gjennomsnitt 90% av intervallene vi lager inneholde denne ukjente (sanne) andelen.

Oppgave 2

X_1 og X_2 er stokastiske variabler med $\mathrm{E}(X_1) = \mathrm{E}(X_2) = 2$, $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1$ og $\mathrm{Cov}(X_1, X_2) = \frac{1}{2}$.

$$
\begin{aligned}
\mathrm{E}(\frac{1}{2}X_1 + \frac{1}{2}X_2) &= \frac{1}{2}\mathrm{E}(X_1) + \frac{1}{2}\mathrm{E}(X_2) = \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 2 = 2 \\
\mathrm{Var}(\frac{1}{2}X_1 + \frac{1}{2}X_2) &= (\frac{1}{2})^2\mathrm{Var}(X_1) + (\frac{1}{2})^2\mathrm{Var}X_2 + 2\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\mathrm{Cov}(X_1, X_2) \\
&= \frac{1}{4}\cdot 1 + \frac{1}{4}\cdot 1 + \frac{2}{4}\cdot\frac{1}{2} = \frac{3}{4} = 0.75
\end{aligned}
$$

Videre, la $X_1, X_2, \ldots, X_{10}$ være stokastiske variabler med $\mathrm{E}(X_i) = 2$ og $\mathrm{Var}(X_i) = 1$ for $i = 1, 2, \ldots, 10$ og videre la $\mathrm{Cov}(X_i, X_j) = \frac{1}{2}$ for alle $i = 1, 2, \ldots, 10$ og $j = 1, 2, \ldots, 10$ slik at $i \neq j$. La $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$.

$$\mathrm{E}(\bar{X}) = \mathrm{E}(\frac{1}{10}\sum_{i=1}^{10} X_i) = \frac{1}{10}\sum_{i=1}^{10}\mathrm{E}(X_i) = \frac{1}{10}\sum_{i=1}^{10} 2 = 2$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{1}{10}\sum_{i=1}^{10} X_i) = (\frac{1}{10})^2\sum_{i=1}^{10}\mathrm{Var}(X_i) + 2\sum_{i=1}^{10}\sum_{j=1}^{i-1}\frac{1}{10}\cdot\frac{1}{10}\mathrm{Cov}(X_i, X_j)$$

$$= (\frac{1}{10})^2\sum_{i=1}^{10} 1 + 2\cdot 45\cdot\frac{1}{10}\cdot\frac{1}{10}\cdot\frac{1}{2} = \frac{1}{10} + \frac{45}{100} = \frac{55}{100} = 0.55$$

Oppgave 3

a)

$$\mathrm{P}(X_1 > 190) = \mathrm{P}\left(\frac{X_1 - 181}{6} > \frac{190 - 181}{6}\right)$$
$$= \mathrm{P}(Z > 1.5) = 1 - \mathrm{P}(Z \leq 1.5) = 1 - 0.9332 = 0.0668.$$

Regn først ut

$$\mathrm{P}(X_1 > 185) = \mathrm{P}\left(\frac{X_1 - 181}{6} > \frac{185 - 181}{6}\right)$$
$$= \mathrm{P}(Z > 0.67) = 1 - \mathrm{P}(Z \leq 0.67) = 1 - 0.7486 = 0.2514.$$

Vi bruker så definisjonen på betinget sannsynlighet

$$\begin{aligned}\mathrm{P}(X_1 > 190 | X_1 > 185) &= \frac{\mathrm{P}(X_1 > 190 \cap X_1 > 185)}{\mathrm{P}(X_1 > 185)} \\ &= \frac{\mathrm{P}(X_1 > 190)}{\mathrm{P}(X_1 > 185)} \\ &= 0.0668/0.2514 \\ &= 0.2657.\end{aligned}$$

Siden X_1 og X_2 er uavhengige, får vi

$$\mathrm{P}(X_1 > 190 | X_2 > 185) = \mathrm{P}(X_1 > 190) = 0.0668.$$

b) En god estimator skal være forventningsrett og ha så liten varians som mulig

Siden utvalgene består av uavhengige identisk fordelte stokastiske variabler er $\mathrm{E}[\bar{X}] = \mu$ og $\mathrm{Var}[\bar{X}] = \sigma^2/n$ og $\mathrm{E}[\bar{Y}] = \mu$ og $\mathrm{Var}[\bar{Y}] = \sigma^2/m$. Videre er

$$\mathrm{E}[\hat{\mu}] = a\mathrm{E}[\bar{X}] + b\mathrm{E}[\bar{Y}] = a\mu + b\mu = (a + b)\mu$$

og, siden $\bar{X}$ og $\bar{Y}$ er uavhengige har vi at

$$\text{Var}[\hat{\mu}] = a^2\text{Var}[\bar{X}] + b^2\text{Var}[\bar{Y}] = a^2\sigma^2/n + b^2\sigma^2/m = \sigma^2\left(\frac{a^2}{n} + \frac{b^2}{m}\right).$$

Siden $\hat{\mu}$ skal være en forventningsrett estimator for μ, blir

$$\begin{aligned} \text{E}[\hat{\mu}] &= \mu \\ (a+b)\mu &= \mu \\ b &= 1-a \end{aligned}$$

og variansen til estimatoren kan skrives som

$$\text{Var}[\hat{\mu}] = \sigma^2\left(\frac{a^2}{n} + \frac{(1-a)^2}{m}\right).$$

Den mest effesiente estimatoren oppnås ved å minimere variansen med hensyn på a. Vi utfører minimeringen ved å sette den deriverte lik 0 og får

$$\begin{aligned} \frac{\text{d}}{\text{d}a}\text{Var}[\hat{\mu}] &= 0 \\ \sigma^2\left(\frac{2a}{n} + \frac{-2(1-a)}{m}\right) &= 0 \\ 2\sigma^2\left(\frac{a}{n} + \frac{a-1}{m}\right) &= 0 \\ a(1/n + 1/m) &= 1/m \\ a &= \frac{n}{n+m} \end{aligned}$$

og $b = m/(n+m)$. Estimatet blir

$$\hat{\mu} = \frac{64}{256}\cdot 180 + \frac{192}{256}\cdot 183 = 182.25$$

c) Null- og alternativ hypotese:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

Siden observasjonene kommer fra en normalfordeling har vi, når nullhypotesen er sann, at

$$T_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n + S_2^2/m}}$$

er omtrent t-fordelt med ν frihetsgrader, der

$$\nu = \frac{(s_1^2/n + s_2^2/m)^2}{(s_1^2/n)^2/(n-1) + (s_2^2/m)^2/(m-1)}.$$

Det betyr at

$$\nu = \frac{(6.0^2/64 + 5.5^2/192)^2}{(6.0^2/64)^2/(64-1) + (5.5^2/192)^2/(192-1)} = 100.6$$

som må rundes ned til $\nu = 100$ frihetsgrader.

La $t_0 = (\bar{x} - \bar{y})/\sqrt{s_1^2/n + s_2^2/m}$, da blir forkastingsområdet til testen: forkast H_0 hvis $t_0 > t_{\alpha/2,\nu}$ eller $t_0 < -t_{\alpha/2,\nu}$, der $t_{0.025,100} = 1.984$.

Observasjonene gir

$$t_0 = (180 - 183)/\sqrt{6.0^2/64 + 5.5^2/192} = -3.5354$$

som ligger i forkastingsområdet og konklusjonen blir forkast H_0. Det er ikke rimelig å anta at de to forskningsgruppene har gjort utvalg fra samme populasjon.

Oppgave 4

a) Det er en klar trend med at økende gestasjonsalder gir økende fødselsvekt og trenden virker lineær. Videre virker det som spredningen av fødselsvekt er omtrent den samme for alle gestasjonsaldere. Det er ikke mulig å forsikre seg om at observasjonene er normalfordelte med så få observasjoner, men det er ingen klare tegn på at antagelsen er brutt. Alt i alt, virker en lineær regresjonsmodell rimelig.

β_0 og β_1 estimeres med minste kvadraters metode ved å velge b_0 og b_1 som minimere summen av kvadratfeil for den estimerte modellen

$$\text{SSE} = \sum_{i=1}^{17}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{17}(y_i - b_0 - b_1 x_i)^2.$$

Dette gjøres ved å sette de deriverte $\partial\text{SSE}/\partial b_0$ og $\partial\text{SSE}/\partial b_1$ lik 0 og løse det lineære ligningsystemet med hensyn på b_0 og b_1.

Den predikerte verdien for $x_0 = 40$ blir

$$\hat{y}_0 = b_0 + b_1 x_0 = -4.02 + 0.18 \cdot 40 = 3.18.$$

b) La $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, da har vi på grunn av uavhengighet mellom $Y_1, Y_2, \ldots, Y_n$ at

$$\begin{aligned}
\text{Var}[B_1] &= \text{Var}\left[\frac{1}{S_{xx}}\sum_{i=1}^n (x_i - \bar{x})Y_i\right] \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}[Y_i] \\
&= \sigma^2 \frac{S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

Dermed er (siden B_1 og $\bar{Y}$ uavhengige)

$$\begin{aligned}
\text{Var}[\hat{Y}_0] &= \text{Var}[B_0 + B_1 x_0] = \text{Var}[\bar{Y} + B_1(x_0 - \bar{x})] \\
&= \text{Var}[\bar{Y}] + (x_0 - \bar{x})^2 \text{Var}[B_1] \\
&= \sigma^2/n + (x_0 - \bar{x})^2 \sigma^2 / S_{xx} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

Det vil si at standardavviket til estimatoren er

$$\sigma\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}.$$

$\hat{Y}_0$ er en estimator for forventningsverdien av fødselsvekt for gestasjonsalder x_0 og kan brukes til å konstruere et konfidensintervall for forventningsverdien. Bredden på konfidensintervallet vil være avhengig av variansen til estimatoren som ble brukt for å konstruere det og uttrykket vi har funnet viser at variansen øker jo større $|x_0-\bar{x}|$ er. Det er derfor mindre usikkerhet for $x_0=39$ enn ved $x_0=29$ fordi $\bar{x}$ er nærmere $x_0=39$ enn $x_0=29$.

Oppgave 5

$$F_Y(y)=\int_{-\infty}^{y} f_Y(t)\mathrm{d}t=\begin{cases}0, & y<0,\\ \int_0^y 1\mathrm{d}t, & 0\le y<1\\ 1, & y\ge 1\end{cases}=\begin{cases}0, & y<0,\\ y, & 0\le y<1,\\ 1, & y\ge 1.\end{cases}$$

Legg først merke til at siden verdien av Y ligger mellom 0 og 1, så kan $-\ln(Y)/\lambda$ bare oppnå verdier større enn 0 Det vil si at for $x>0$ kan vi bruke

$$F_X(x)=\mathrm{P}(X\le x)=\mathrm{P}(-\ln(Y)/\lambda\le x)=\mathrm{P}(\ln(Y)\ge-\lambda x)=1-\mathrm{P}(Y\le \mathrm{e}^{-\lambda x}),$$

der ulikhetstegnet må snus fordi det ganges med -1 på begge sider av ulikheten. Dette betyr at

$$F_X(x)=1-F_Y(\mathrm{e}^{-\lambda x})=1-\mathrm{e}^{-\lambda x},\quad \text{for } x>0.$$

X kan ikke bli mindre enn 0 så

$$F_X(x)=\begin{cases}0, & x\le 0\\ 1-\mathrm{e}^{-\lambda x}, & x>0\end{cases}.$$

Sannsynlighetsfordeling er

$$f_X(x)=\frac{\mathrm{d}}{\mathrm{d}x}F_X(x)=\lambda\mathrm{e}^{-\lambda x},\quad x>0.$$

Dette er en eksponentialfordeling med forventningsverdi $1/\lambda$.