TMA4240/4245
Statistikk
Eksamen august 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a) Ved kast av to terninger er det 36 mulige utfall: $(1,1)$, ..., $(6,6)$.

La $Y_2 = X_1 + X_2$.

$P(Y_2 = 12) = 1/36$.

Det er 6 gunstige utfall med sum 10 eller større: $(4,6)$, $(6,4)$, $(5,6)$, $(6,5)$, $(5,5)$, $(6,6)$. $P(Y_2 \geq 10) = \frac{6}{36} = \frac{1}{6}$.

$E(Y_2) = E(X_1) + E(X_2) = 2E(X_1) = 2 \cdot \frac{1}{6} \sum_{l=1}^{6} l = \frac{2}{6}\frac{7 \cdot 6}{2} = 7$.

b) $E(Z_1) = \frac{1}{6} \sum_{l=1}^{5} l + \frac{1}{6} \cdot 0 = \frac{1}{6}\frac{6 \cdot 5}{2} + 0 = 2.5$.

$E(Z_k | Z_{k-1} = 30) = \frac{1}{6} \sum_{l=1}^{5} (30 + l) + \frac{1}{6} \cdot 0 = \frac{1}{6}(5 \cdot 30 + \frac{6 \cdot 5}{2} + 0) = 25 + 2.5 = 27.5$.

Anta at en person har score s etter kast $k-1$. Forventet score etter neste kast er da (fra forrige svar): $E(Z_k | Z_{k-1} = s) = s \cdot \frac{5}{6} + 2.5$. Dersom denne forventningen er større enn nåværende score s, så lønner det seg å fortsette kasting.

$s \cdot \frac{5}{6} + 2.5 > s$, $2.5 > \frac{s}{6}$, $15 > s$

Oppgave 2

a) La X være pris i Euro. Da er $Y = 9.2X$ pris i kroner. En lineærkombinasjon av normalfordelte variable er normalfordelt. $E(Y) = 9.2E(X) = 9.2 \cdot 100 = 920$. $Var(Y) = 9.2^2 Var(X) = 9.2^2 \cdot 20^2 = 184^2$.

$P(Y > 1000) = P(Z > \frac{1000-920}{\sqrt{184}}) = P(Z > 0.43) = 1 - P(Z < 0.43) = 0.33$

$P(Y > 1100 | Y > 1000) = \frac{P(Y>1100 \cap Y>1000)}{P(Y>1000)} = \frac{P(Y>1100)}{0.33}$

Her er: $P(Y > 1100) = P(Z > \frac{1100-920}{184}) = P(Z > 0.97) = 0.16$

Da er $P(Y > 1100 | Y > 1000) = \frac{0.16}{0.33} = 0.49$

b) H_0: $\mu = 100$, H_1: $\mu > 100$.

$\bar{X} \sim N(\mu, 20^2/n)$, og her er $n = 20$.

Forkaster H_0 dersom $Z = \frac{\bar{X}-100}{20/\sqrt{20}} > 1.645$.

Vi observerer $z = \frac{120-100}{\sqrt{20}} = \sqrt{20} = 4.5$.

Da forkastes H_0.

c) Teststyrken er sannsynligheten for å forkaste H_0 gitt at H_1 er sann ($\mu = 110$) i dette tilfellet:

$$P(\frac{\bar{X} - 100}{20/\sqrt{20}} > 1.64|\mu = 110) = P(\frac{\bar{X} - 110 + 110 - 100}{20/\sqrt{20}} > 1.645|\mu = 110)$$

$$= P(Z > 1.645 - \frac{110 - 100}{20/\sqrt{20}}) = P(Z > -0.59) = 0.72$$

Kravet er at teststyrken skal være minst 0.95. $P(\frac{\bar{X}-100}{20/\sqrt{n}} > 1.64|\mu = 110) = 0.95$

$P(Z > 1.645 - \frac{110-100}{20/\sqrt{n}}) = 0.95.$

Kravet for n blir da:

$1.645 - \frac{110-100}{20/\sqrt{n}} = -1.645$

$2 \cdot 1.645 = \sqrt{n}/2$

$n = (2 \cdot 2 \cdot 1.645)^2 = 43.43.$

Antall data n må være minst 44 for å oppnå så høy teststyrke.

Oppgave 3

a) Et normalsannsynlighetsplott benyttes til å vurdere om det er rimelig å anta at et gitt datasett er normalfordelt. Dersom punktene i normalsannsynlighetsplottet ligger tilnærmet på et rett linje er det rimelig å anta at observasjonene er fra en normalfordeling.

I residualplottet sees ingen spesielle strukturer. Spesielt ser man ingen spesielle sammenhanger mellom x og estimert residual, og variansen til estimert residual synes ikke å variere med x. Punktene i normalsannsynlighetsplottet ligger tilnærmet på en rett linje. At punktene for de minste og største x-verdiene avviker noe fra den rette linja er som man kan forvente i et slikt plott. Det er dermed ikke noe i de to plottene som indikerer at den spesifiserte modellen ikke er rimelig for dette datasettet.

b) Man skal foretrekke en estimator som er forventingsrett. Hvis flere estimatorer er forventingsrette skal man foretrekke den som har minst varians.

Starter med å sjekke om $\widetilde{b}$ og $\widehat{b}$ er forventningsrette. Ved å benytte regneregler for forventingsverdi og antagelsen $\mathrm{E}[Y_i] = bx_i$ får vi

$$\begin{aligned} \mathrm{E}\left[\widetilde{b}\right] &= \mathrm{E}\left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right] = \frac{\mathrm{E}\left[\sum_{i=1}^n Y_i\right]}{\sum_{i=1}^n x_i} \\ &= \frac{\sum_{i=1}^n \mathrm{E}\left[Y_i\right]}{\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n bx_i}{\sum_{i=1}^n x_i} = \frac{b\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i} = b, \end{aligned}$$

og

$$\begin{aligned} \mathrm{E}\left[\widehat{b}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n Y_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^n Y_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[Y_i\right] = \frac{1}{n}\sum_{i=1}^n bx_i = \frac{b}{n}\sum_{i=1}^n x_i = b\bar{x}. \end{aligned}$$

Vi ser at $\widetilde{b}$ er forventingsrett, mens $\widehat{b}$ er forventingsskjev. Siden det er oppgitt at $\widehat{\widehat{b}}$ er forventningsrett skal vi derfor foretrekke $\widetilde{b}$ eller $\widehat{\widehat{b}}$. For å velge mellom disse må vi regne ut variansen til $\widetilde{b}$. Ved å benytte regneregler for varians, at Y_i-ene er uavhengige og at $\text{Var}[Y_i] = \sigma^2$ for alle $i = 1, 2, \ldots, n$ får vi

$$\begin{aligned}
\text{Var}\left[\widetilde{b}\right] &= \text{Var}\left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right] = \frac{1}{\left(\sum_{i=1}^n x_i\right)^2}\text{Var}\left[\sum_{i=1}^n Y_i\right] \\
&= \frac{1}{\left(\sum_{i=1}^n x_i\right)^2}\sum_{i=1}^n \text{Var}\left[Y_i\right] = \frac{1}{\left(\sum_{i=1}^n x_i\right)^2}\sum_{i=1}^n \sigma^2 \\
&= \frac{n\sigma^2}{\left(\sum_{i=1}^n x_i\right)^2}.
\end{aligned}$$

Setter vi inn oppgitte tall for n og x_i-ene, og for σ^2 får vi følgende varianser for de to forventningsrette estimatorene,

$$\text{Var}\left[\widetilde{b}\right] = \frac{10 \cdot 2^2}{36.5^2} = 0.030, \quad , \quad \text{Var}\left[\widehat{\widehat{b}}\right] = \frac{2^2}{152.25} = 0.026.$$

Siden $\text{Var}\left[\widehat{\widehat{b}}\right] < \text{Var}\left[\widetilde{b}\right]$ er estimatoren $\widehat{\widehat{b}}$ å foretrekke.

c) Rimelighetsfunksjonen for b blir her

$$\begin{aligned}
\underline{\underline{L(b)}} &= f(y_1, y_2, \ldots, y_n; b) = \prod_{i=1}^n f(y_i; b) \\
&= \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_i - bx_i)^2\right\}\right] \\
&= \underline{\underline{\frac{1}{\left(\sqrt{2\pi}\right)^n}\frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - bx_i)^2\right\}}}.
\end{aligned}$$

For å finne SME danner vi først log-rimelighetsfunksjonen og beregner deretter den derivere av denne med hensyn på b,

$$\begin{aligned}
l(b) &= \ln L(b) = -n\ln\left(\sqrt{2\pi}\right) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - bx_i)^2 \\
l'(b) &= -\frac{1}{2\sigma^2}\sum_{i=1}^n \left[2\left(y_i - bx_i\right)\cdot(-x_i)\right] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i y_i - bx_i^2) = \frac{1}{\sigma^2}\left[\sum_{i=1}^n x_i y_i - b\sum_{i=1}^n x_i^2\right].
\end{aligned}$$

Bestemmer så for hvilken verdi av b rimelighetsfunksjonen $l(b)$ har sitt maksimum ved å løse $l'(b) = 0$ med hensyn på b,

$$l'(b) = 0 \Rightarrow \sum_{i=1}^n xiy_i = b\sum_{i=1}^n x_i^2 \Rightarrow b = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

SME for b er dermed

$$\underline{\underline{\widehat{\widehat{b}} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.}}$$

d) Estimatoren $\widehat{\widehat{b}}$ er en lineærkombinasjon av Y_i-ene som er antatt uavhengige og normalfordelte. Siden vi vet at enhver lineærkombinasjon av uavhengige normalfordelte variabler er normalfordelt, er derfor $\widehat{\widehat{b}}$ normalfordelt.

For å utlede et $(1-\alpha)\cdot 100\%$ konfidensintervall for b tar vi utgangspunkt i en standardisert versjon av $\widehat{\widehat{b}}$ som vi vet er standard-normalfordelt,

$$Z = \frac{\widehat{\widehat{b}} - b}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}} \sim n(z;0,1).$$

Vi får dermed at

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\widehat{b}} - b}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Løser så hver av de to ulikhetene over med hensyn på b. Venstre ulikhet gir

$$-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\widehat{b}} - b}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}} \quad \Leftrightarrow \quad -\widehat{\widehat{b}} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}} \leq -b \quad \Leftrightarrow \quad \widehat{\widehat{b}} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}} \geq b,$$

og høyre ulikhet gir

$$\frac{\widehat{\widehat{b}} - b}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}} \leq z_{\frac{\alpha}{2}} \quad \Leftrightarrow \quad -b \leq -\widehat{\widehat{b}} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}} \quad \Leftrightarrow \quad b \geq \widehat{\widehat{b}} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}$$

Hvis vi så skriver disse to ulikhetene sammen igjen med b i midten får vi at

$$P\left(\widehat{\widehat{b}} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}} \leq b \leq \widehat{\widehat{b}} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}\right) = 1-\alpha.$$

Et $(1-\alpha)\cdot 100\%$ konfidensintervall for b er dermed

$$\underline{\underline{\left[\widehat{\widehat{b}} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}, \widehat{\widehat{b}} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}\right]}}$$

Med $\alpha = 0.10$ finner vi i tabell at $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$, og innsatt observerte tall får vi

$$\widehat{\widehat{b}} = \frac{331.65}{152.25} = 2.178325.$$

Et 90% konfidensintervall for b blir dermed

$$\left[\frac{331.65}{152.25} - 1.645\sqrt{\frac{2^2}{152.25}}, \frac{331.65}{152.25} - 1.645\sqrt{\frac{2^2}{152.25}}\right] = \underline{\underline{[1.9117, 2.4450]}}$$

e) La x_0 være dosen med medisin som gis til denne nye pasienten og la Y_0 være tilhørende målt konsentrasjon etter et døgn. Kravet som er verbalt formulert i oppgaven kan da matematisk uttrykkes som at man ønsker å bestemme den høyeste dosen x_0 slik at

$$P\left(Y_0 \leq 10\right) \geq 0.95. \tag{3.1}$$

Vi vet at $Y_0 \sim n(y_0; bx_0, \sigma)$, men man må huske at verdien til b er ukjent slik at b ikke kan inngå i svaret.

La derfor, tilsvarende som det er vanlig å gjøre når man utleder prediksjonsintervaller,

$$\widehat{\widehat{Y}}_0 = \widehat{\widehat{b}} x_0$$

være predikert verdi for Y_0. Vi vil da ha at $Y_0 - \widehat{\widehat{Y}}_0 = Y_0 - \widehat{\widehat{b}} x_0$ er normalfordelt fordi det er en lineærkombinasjon av de uavhengige og normalfordelte variablene Y_0 og $\widehat{\widehat{b}}$. Forventningsverdi og varians for $Y_0 - \widehat{\widehat{Y}}_0$ blir

$$\begin{aligned}
\mathrm{E}\left[Y_0 - \widehat{\widehat{b}} x_0\right] &= \mathrm{E}\left[Y_0\right] - \mathrm{E}\left[\widehat{\widehat{b}}\right] x_0 = bx_0 - bx_0 = 0, \\
\mathrm{Var}\left[Y_0 - \widehat{\widehat{b}} x_0\right] &= \mathrm{Var}\left[Y_0\right] + \mathrm{Var}\left[\widehat{\widehat{b}}\right] \left(-x_0\right)^2 \\
&= \sigma^2 + \frac{\sigma^2}{\sum_{i=1}^n x_i^2} \cdot x_0^2 = \sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)
\end{aligned}$$

Dermed har man at

$$Z = \frac{Y_0 - \widehat{\widehat{b}} x_0}{\sqrt{\sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}} \sim n(z; 0, 1).$$

Vi får dermed at

$$P(Z \leq z_{0.05}) = P\left(\frac{Y_0 - \widehat{\widehat{b}} x_0}{\sqrt{\sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}} \leq z_{0.05}\right) = 0.95$$

og ved å løse ulikheten her med hensyn på Y_0 får vi også

$$P\left(Y_0 \leq \widehat{\widehat{b}} x_0 + z_{0.05} \sqrt{\sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}\right) = 0.95.$$

Ved å kreve at høyresiden i ulikheten er lik 10 får vi oppfylt $P(Y_0 \leq 10) = 0.95$. Man skal merke seg at dette kravet innebærer at den nye dosen som skal brukes, x_0, blir en funksjon av estimatoren $\widehat{\widehat{b}}$ og dermed en stokastisk variabel. Den nye dosen som skal settes er dermed gitt som løsning av ligningen

$$\widehat{\widehat{b}} x_0 + z_{0.05} \sqrt{\sigma^2 \left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)} = 10.$$

Å gi en generell løsning av denne ligningen med hensyn på x_0 er vanskelig, men for oss er det tilstrekkelig å løse den for observert verdi av $\widehat{\widehat{b}}$. Heretter lar vi dermed $\widehat{\widehat{b}}$ betegne observert verdi for estimatoren, dvs $\widehat{\widehat{b}} = 331.65/152.25 = 2.178325$. Vi får kravet

$$10 - \widehat{\widehat{b}}x_0 = z_{0.05}\sqrt{\sigma^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}.$$

Kvadrerer på begge sider for å bli kvitt kvadratrottegnet (men må huske på at vi da kan introdusere nye løsninger som har $10 - \widehat{\widehat{b}}x_0 < 0$),

$$\begin{aligned}(10 - \widehat{\widehat{b}}x_0)^2 &= z_{0.05}^2\sigma^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right), \\ 10^2 - 2 \cdot 10 \cdot \widehat{\widehat{b}}x_0 + \widehat{\widehat{b}}^2 x_0^2 &= z_{0.05}^2\sigma^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right).\end{aligned}$$

Hvis vi setter inn observerte verdier for $\widehat{\widehat{b}}$ og $\sum_{i=1}^n x_i^2$, verdier for σ^2 og $z_{0.05}$, og samler konstanter, lineære og kvadratiske ledd i x_0 får vi kravet

$$89.1759 - 43.5665x_0 + 4.6740x_0^2 = 0. \tag{3.2}$$

Løser man denne andregradsligningen får man løsningene

$$x_0 = 6.2857 \quad \text{og} \quad x_0 = 3.0353.$$

For at disse skal være løsning på vår opprinnelig ligning må vi ha at $10 - \widehat{\widehat{b}}x_0 > 0$, noe $x_0 = 6.2857$ ikke gjør. Man bør følgelig gi den nye pasienten dosen

$$\underline{\underline{x_0 = 3.0353}}.$$