TMA4245 Statistikk
Eksamen august 2014

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

a)

$$\mathrm{P}(Y > 53) = 1 - \mathrm{P}(Y \leq 53) = 1 - \mathrm{P}\left(\frac{Y-50}{2} \leq \frac{53-50}{2}\right) = 1 - \Phi(1.50) = 1 - 0.9332 = \underline{\underline{0.0668}}.$$

Vekten av ei fylt flaske, $X + Y$, er en lineærkombinasjon av uavhengige normalfordelte variabler og dermed selv normalfordelt. Dessuten har vi at

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 510 + 50 = 560$$

og

$$\mathrm{Var}[X + Y] = \mathrm{E}[X] + \mathrm{Var}[Y] = 12^2 + 2^2 = 148.$$

Dermed er $X + Y$ normalfordelt med forventningsverdi lik 560 gram av standardavvik lik $\sqrt{148}$ gram.

La R_1 og R_2 være vektene av to vilkårlig valgte flasker fylt med punlver. Vi vet at R_1 og R_2 er uavhengige og at hver av dem er normalfordelt med forventningsverdi lik 560 gram og standardavvik lik $\sqrt{148}$ gram. Vektforskjellen mellom to vilkårlig valgte flasker med pulver blir $S = R_1 - R_2$. Siden S er en lineærkombinasjon av uavhengige normalfordelte variabler blir også S normalfordelt. Får dessuten at

$$\mathrm{E}[S] = \mathrm{E}[R_1 - R_2] = \mathrm{E}[R_1] - \mathrm{E}[R_2] = 560 - 560 = 0$$

og siden R_1 og R_2 er uavhengige

$$\mathrm{Var}[S] = \mathrm{Var}[R_1 - R_2] = \mathrm{Var}[R_1] + (-1)^2\mathrm{Var}[R_2] = 148 + 148 = 296.$$

Dermed, siden fordelingen til S er symmetrisk om null, får man at

$$\mathrm{P}(|S| > 15) = 2\mathrm{P}(S > 15) = 2\mathrm{P}\left(\frac{S-0}{\sqrt{296}} > \frac{15-0}{\sqrt{296}}\right)$$

$$= 2\left(1 - \mathrm{P}\left(\frac{S-0}{\sqrt{296}} > \frac{15-0}{\sqrt{296}}\right)\right) = 2(1 - \Phi(0.87)) = 2(1 - 0.8078) = \underline{\underline{0.3855}}.$$

b) La $R = X + Y$ være vekten av ei flaske fylt med pulver. Da får man at

$$\mathrm{P}(R < 540) = \mathrm{P}\left(\frac{X+Y-560}{\sqrt{148}} < \frac{540-560}{\sqrt{148}}\right) = \Phi(-1.64) = \underline{\underline{0.0505}}.$$

Man har at $\underline{\underline{U \sim b(u; 24, p)}}$ der $p = \mathrm{P}(R < 540) = 0.0505$ fordi

- man har 24 uavhengige forsøk,
- hvert forøk gir suksess (undervektig flaske) eller fiasko (ikke undervektig flaske),
- sannsynligheten for suksess er lik i alle forsøkene, og
- U er antall suksesser.

$$\mathrm{P}(U \geq 1) = 1 - P(U < 1) = 1 - P(U = 0)1 - (1-p)^{24} = 1 - (1 - 0.0505)^{24} = \underline{\underline{0.7117}}.$$

c) Man krever at $\mathrm{P}(R < 540) \leq 0.01$ når man har at $R \sim n(r, \mu + 50, \sqrt{148})$. Dette gir at

$$\mathrm{P}\left(\frac{R - (\mu + 50)}{\sqrt{148}} < \frac{540 - (\mu + 50)}{\sqrt{148}}\right) = \Phi\left(\frac{490 - \mu}{\sqrt{148}}\right) \leq 0.01$$

som er oppfylt dersom

$$\frac{490 - \mu}{\sqrt{148}} \leq -z_{0.01} = -2.326 \Rightarrow -\mu \leq -2.326\sqrt{148} - 490$$

$$\Rightarrow \mu \geq 490 + 2.326\sqrt{148} = \underline{\underline{518.30}}.$$

La V betegne antall undervektige flasker i 50 kartonger. Ut fra samme argumentasjon som for U vil også V være binomisk fordelt, men nå med $50 \cdot 24 = 1200$ forsøk og sannsynlighet 0.01 for suksess. Dermed får man at

$$\mathrm{E}[V] = 1200 \cdot 0.01 = \underline{\underline{12}}.$$

d) $\widehat{\mu}$ er en lineærkombinasjon av uavhengige normalfordelte variabler og blir derfor selv normalfordelt. Dessuten får man at

$$\mathrm{E}\left[\widehat{\mu}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\mu = \frac{1}{n}\cdot n\mu = \mu,$$

og siden X_i'ene er uavhengige

$$\mathrm{Var}\left[\widehat{\mu}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] =$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} 12^2 = \frac{1}{n^2}\cdot n \cdot 12^2 = \frac{12^2}{n}.$$

Dermed får man at

$$Z = \frac{\widehat{\mu} - \mu}{\sqrt{\frac{12^2}{n}}} \sim n(z; 0, 1)$$

slik at

$$\mathrm{P}\left(-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\mu} - \mu}{\sqrt{\frac{12^2}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Må løse de to ulikhetene hver for seg. Den første ulikheten gir

$$-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\mu}-\mu}{\sqrt{\frac{12^2}{n}}} \Rightarrow \mu \leq \widehat{\mu} + z_{\frac{\alpha}{2}}\sqrt{\frac{12^2}{n}},$$

og den andre ulikheten gir

$$\frac{\widehat{\mu}-\mu}{\sqrt{\frac{12^2}{n}}} \leq z_{\frac{\alpha}{2}} \Rightarrow \widehat{\mu} - z_{\frac{\alpha}{2}}\sqrt{\frac{12^2}{n}} \leq \mu.$$

Vi får dermed at

$$\mathrm{P}\left(\widehat{\mu} - z_{\frac{\alpha}{2}}\sqrt{\frac{12^2}{n}} \leq \mu \leq \widehat{\mu} + z_{\frac{\alpha}{2}}\sqrt{\frac{12^2}{n}}\right) = 1-\alpha,$$

slik at et $(1-\alpha)\cdot 100\%$-konfidensintervall for μ blir

$$\underline{\underline{\left[\widehat{\mu} - z_{\frac{\alpha}{2}}\sqrt{\frac{12^2}{n}}, \widehat{\mu} + z_{\frac{\alpha}{2}}\sqrt{\frac{12^2}{n}}\right]}}.$$

Innsatt tall får vi $n = 24$, $\alpha = 0.05$, $z_{\frac{\alpha}{2}} = 1.96$, $\widehat{\mu} = \bar{x} = 513.4$, og dermed intervallet

$$\left[513.4 - 1.96\sqrt{\frac{12^2}{24}}, 513.4 + 1.96\sqrt{\frac{12^2}{24}}\right] = \underline{\underline{[508.60, 518.20]}}.$$

Som funksjon av n får man at lengden på konfidensintervallet blir

$$L = 2\cdot 1.96\sqrt{\frac{12^2}{n}}.$$

Kravet $L < 8$ gir dermed at

$$2\cdot 1.96\sqrt{\frac{12^2}{n}} < 8 \Rightarrow \frac{1.96\cdot 12}{4} < \sqrt{n} \Rightarrow n > \left(\frac{1.96\cdot 12}{4}\right)^2 = 34.57.$$

For at lengden på konfidensintervallet skal bli kortere enn 8 gram må man dermed ha $\underline{\underline{n \geq 35}}$.

e) Man har nå at $V_i \sim n(v_i; \beta\mu_i, 12)$. Siden V_i'ene er antatt uavhengige får man at rimelighetsfunksjonen blir

$$L(\beta) = \prod_{i=1}^{k} n(v_i; \beta\mu_i, 12) = \prod_{i=1}^{k}\left[\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{12}\exp\left\{-\frac{1}{2\cdot 12^2}(v_i-\beta\mu_i)^2\right\}\right].$$

Log-rimelighetsfunksjonen blir dermed

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{k}\left[-\frac{1}{2}\ln(2\pi) - \ln(12) - \frac{1}{2\cdot 12^2}(v_i-\beta\mu_i)^2\right]$$

$$= -\frac{k}{2}\ln(2\pi) - k\ln(12) - \frac{1}{2\cdot 12^2}\sum_{i=1}^{k}(v_i - \beta\mu_i)^2.$$

Deriverer for å finne maksimum,

$$l'(\beta) = -\frac{1}{2\cdot 12^2}\sum_{i=1}^{k} 2(v_i - \beta\mu_i)(-\mu_i) = \frac{1}{12^2}\left[\sum_{i=1}^{k} v_i\mu_i - \beta\sum_{i=1}^{k}\mu_i^2\right].$$

Dermed får vi at

$$l'(\beta) = 0 \Rightarrow \beta = \frac{\sum_{i=1}^{k} v_i\mu_i}{\sum_{i=1}^{k}\mu_i^2}.$$

Sannsynlighetsmaksimeringsestimatoren blir dermed

$$\underline{\underline{\widehat{\beta} = \frac{\sum_{i=1}^{k} V_i\mu_i}{\sum_{i=1}^{k}\mu_i^2}}}.$$

Innsatt tall får vi

$$\widehat{\beta} = \frac{2440526}{2462800} = \underline{\underline{0.9910}}.$$

Oppgave 2

a)

$\mathrm{P}(X > Y) = \mathrm{P}(X = 1, Y = 0) + \mathrm{P}(X = 2, Y = 0) + \mathrm{P}(X = 2, Y = 1) = 0.06 + 0.04 + 0.10 = \underline{\underline{0.2}}$.

Marginal punktsannsynlighet for X blir

$$\begin{aligned} g(0) &= \mathrm{P}(X = 0) = 0.10 + 0.25 + 0.15 = 0.5, \\ g(1) &= \mathrm{P}(X = 1) = 0.06 + 0.15 + 0.09 = 0.3, \\ g(2) &= \mathrm{P}(X = 2) = 0.04 + 0.10 + 0.06 = 0.2. \end{aligned}$$

Marginal punktsannsynlighet for Y blir

$$\begin{aligned} h(0) &= \mathrm{P}(Y = 0) = 0.10 + 0.06 + 0.04 = 0.2, \\ h(1) &= \mathrm{P}(Y = 1) = 0.25 + 0.15 + 0.10 = 0.5, \\ h(2) &= \mathrm{P}(Y = 2) = 0.15 + 0.09 + 0.06 = 0.3. \end{aligned}$$

X og Y er uavhengige hvis og bare hvis $f(x, y) = g(x)h(y)$ for alle x og y. Regner ut $g(x)h(y)$ for $x, y = 0, 1, 2$ og får

$x\backslash y$	0	1	2
0	0.10	0.25	0.15
1	0.06	0.15	0.09
2	0.04	0.10	0.06

Vi ser dermed at kravet til uavhengighet er oppfylt, så $\underline{\underline{X \text{ og } Y \text{ er uavhengige}}}$.

Oppgave 3

a)

$$\mathrm{E}\left[\widetilde{\beta}\right] = \mathrm{E}\left[\frac{\sum_{i=1}^n Y_i x_i^2}{\sum_{i=1}^n x_i^4}\right] = \frac{1}{\sum_{i=1}^n x_i^4}\mathrm{E}\left[\sum_{i=1}^n Y_i x_i^2\right] = \frac{1}{\sum_{i=1}^n x_i^4}\sum_{i=1}^n \mathrm{E}\left[Y_i x_i^2\right]$$

$$= \frac{1}{\sum_{i=1}^n x_i^4}\sum_{i=1}^n x_i^2 \mathrm{E}\left[Y_i\right] = \frac{1}{\sum_{i=1}^n x_i^4}\sum_{i=1}^n x_i^2 \beta x_i^2 = \frac{\beta\sum_{i=1}^n x_i^4}{\sum_{i=1}^n x_i^4} = \underline{\underline{\beta}}$$

$$\mathrm{Var}\left[\widetilde{\beta}\right] = \mathrm{Var}\left[\frac{\sum_{i=1}^n Y_i x_i^2}{\sum_{i=1}^n x_i^4}\right] = \left(\frac{1}{\sum_{i=1}^n x_i^4}\right)^2 \mathrm{Var}\left[\sum_{i=1}^n Y_i x_i^2\right] = \frac{1}{\left(\sum_{i=1}^n x_i^4\right)^2}\sum_{i=1}^n \mathrm{Var}\left[Y_i x_i^2\right]$$

$$= \frac{1}{\left(\sum_{i=1}^n x_i^4\right)^2}\sum_{i=1}^n x_i^4 \mathrm{Var}\left[Y_i\right] = \frac{1}{\left(\sum_{i=1}^n x_i^4\right)^2}\sum_{i=1}^n x_i^4(\sigma x_i)^2 = \underline{\underline{\sigma^2 \frac{\sum_{i=1}^n x_i^6}{\left(\sum_{i=1}^n x_i^4\right)^2}}}$$

Begge estimatorene er forventingsrette. Vi vil derfor foretrekke den estimatoren som har minst varians. Innsatt de benyttede verdiene for x_i'ene blir variansene

$$\mathrm{Var}\left[\widehat{\beta}\right] = \sigma^2 \frac{1}{\sum_{i=1}^n x_i^2} = \sigma^2 \frac{1}{82750} = 1.21 \cdot 10^{-5}\sigma^2$$

og

$$\mathrm{Var}\left[\widetilde{\beta}\right] = \sigma^2 \frac{\sum_{i=1}^n x_i^6}{\left(\sum_{i=1}^n x_i^4\right)^2} = \sigma^2 \frac{4.7222 \cdot 10^{12}}{(573216250)^2} = 1.44 \cdot 10^{-5}\sigma^2.$$

Siden $\mathrm{Var}\left[\widehat{\beta}\right] < \mathrm{Var}\left[\widetilde{\beta}\right]$ vil vi dermed <u>foretrekke estimatoren $\widehat{\beta}$</u>.

b) 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 normafordeling.

I residualplottet i (b) sees ingen spesielle strukturer. Spesielt ser man ingen spesielle sammenhenger mellom x og $\widehat{\varepsilon}$ og variansen til $\widehat{\varepsilon}$ synes ikke å variere med x. Punktene i normalsannsynlighetsplottet i (c) ligger tilnærmet på en rett linje. At punktene for de minste og de 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.

c) Vi har at $\widehat{\beta}$ er normalfordelt med forventning og varians som oppgitt i oppgaveteksten. For å utlede et prediksjonsintervall er det naturlig å betrakte $Y_0 - \widehat{\beta}x_0^2$. Siden Y_0 også er normalfordelt og uavhengig av $\widehat{\beta}$ blir også $Y_0 - \widehat{\beta}x_0^2$ normalfordelt siden dette er en lineærkombinasjon av uavhengige normalfordelte variabler. Dessuten får man at

$$\mathrm{E}\left[Y_0 - \widehat{\beta}x_0^2\right] = \mathrm{E}[Y_0] - \mathrm{E}\left[\widehat{\beta}\right]x_0^2 = \beta x_0^2 - \beta x_0^2 = 0$$

og

$$\mathrm{Var}\left[Y_0 - \widehat{\beta}x_0^2\right] = \mathrm{Var}[Y_0] + x_0^4 \mathrm{Var}\left[\widehat{\beta}\right] = \sigma^2 x_0^2 + \frac{\sigma^2 x_0^4}{\sum_{i=1}^n x_i^2} = \sigma^2\left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2}\right).$$

Dermed får vi at

$$Z = \frac{Y_0 - \widehat{\beta} x_0^2}{\sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)}} \sim n(z; 0, 1).$$

Dette gir at

$$\mathrm{P}\left(-z_{\frac{\alpha}{2}} \leq \frac{Y_0 - \widehat{\beta} x_0^2}{\sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)}} \leq z_{\frac{\alpha}{2}} \right) = 1 - \alpha.$$

Må løse hver av de to ulikhetene hver for seg. Den første ulikheten gir

$$-z_{\frac{\alpha}{2}} \leq \frac{Y_0 - \widehat{\beta} x_0^2}{\sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)}} \Rightarrow \widehat{\beta} x_0^2 - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)} \leq Y_0$$

og den andre ulikheten gir

$$\frac{Y_0 - \widehat{\beta} x_0^2}{\sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)}} \leq z_{\frac{\alpha}{2}} \Rightarrow Y_0 \leq \widehat{\beta} x_0^2 + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)}.$$

Dermed får vi at

$$\mathrm{P}\left(\widehat{\beta} x_0^2 - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)} \leq Y_0 \leq \widehat{\beta} x_0^2 + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)} \right) = 1 - \alpha,$$

og prediksjonsintervallet blir dermed

$$\underline{\underline{\left[\widehat{\beta} x_0^2 - z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)}, \widehat{\beta} x_0^2 + z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left(x_0^2 + \frac{x_0^4}{\sum_{i=1}^n x_i^2} \right)} \right]}}.$$

For at prediksjonsintervallet skal bli kort må man velge store verdier for alle x_i'ene. Dette er rimelig siden man i modellen vet at $\mathrm{E}[Y] = 0$ for $x = 0$. Det kan dog nevnes at dersom man velger alle x_i'ene store vil man ikke kunne benytte observerte data til å verifisere modellantagelsene.

d) De to hypotesene blir her

$$H_0 : \beta = \beta_0 = 0.0053 \quad \text{mot} \quad H_1 : \beta < \beta_0$$

Dersom σ^2 hadde hatt en kjent verdi ville man ha benyttet

$$Z = \frac{\widehat{\beta} - \beta_0}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}}$$

som testobservator. Under H_0 vil vi ha at $Z \sim n(z; 0, 1)$. Siden verdien til σ^2 er ukjent erstatter vi σ^2 i uttrykket for Z med dens estimator, $\widehat{\sigma}^2$. Vi får dermed testobservatoren

$$T = \frac{\widehat{\beta} - \beta_0}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n x_i^2}}}.$$

For lettere å se hvilken fordeling denne har, merk at vi alternativt kan skrive T som

$$T = \frac{\frac{\widehat{\beta}-\beta_0}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}}}{\sqrt{\frac{\frac{n-1}{\sigma^2}\widehat{\sigma}^2}{n-1}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}},$$

der $V = (n-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$ er uavhengig av Z. Vi kan dermed konkludere at T er Student t-fordelt med $n-1$ frihetsgrader.

Siden T vil tendere til å bli liten hvis β er liten er det rimelig å forkaste H_0 dersom T er liten nok, dvs. forkaster H_0 dersom $T < k$, der k bestemmes fra kravet

$$\mathrm{P}(T < k|H_0) = \alpha.$$

Dette gir $k = -t_{\alpha,n-1}$, dvs. forkaster H_0 dersom $T < -t_{\alpha,n-1}$.

Innsatt tall får vi $t_{\alpha,n-1} = t_{0.05,19} = 1.729$,

$$\widehat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2} = \frac{n\bar{y}}{\sum_{i=1}^n x_i^2} = \frac{20 \cdot 21.022}{82750} = 0.00508085,$$

og

$$\widehat{\sigma}^2 = \frac{1}{20-1}\left[\sum_{i=1}^n \left(\frac{y_i}{x_i}\right)^2 - 2\widehat{\beta}\sum_{i=1}^n y_i + \widehat{\beta}^2 \sum_{i=1}^n x_i^2\right]$$

$$= \frac{1}{19}\left[2.1656 - 2 \cdot 0.00508085 \cdot 20 \cdot 21.022 + (0.00508085)^2 \cdot 82750\right] = 0.0015495$$

slik at observert verdi for testobservatoren blir

$$t = \frac{0.00508085 - 0.0053}{\sqrt{\frac{0.0015495}{82750}}} = -1.60151.$$

Vi har dermed $t \not< -t_{0.05,19}$, så det er ikke grunnlag for å forkaste H_0.