TMA4240 Statistikk

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsforslag - Eksamen desember 2003

Oppgave 1

Oppgitt: $\int_0^\infty x^r e^{-ax}\mathrm{d}x = \frac{r!}{a^{r+1}}$ for $a > 0$, $r \geq 0$ heltall.

a) For at $g(x)$ skal være en sannsynlighetstetthet, må vi ha $\int_{-\infty}^{\infty} g(x)\mathrm{d}x = 1$, dvs at total sannsynlighet er 1. (Bruker formelen med $r = 1$ og $a = 2$.)

$$\int_{-\infty}^{\infty} g(x)\mathrm{d}x = \int_0^\infty kxe^{-2x}\mathrm{d}x = k \cdot \frac{1}{2^2} = \frac{k}{4}.$$

$k/4 = 1$ gir $k = 4$.

For forventet forsinkelse brukes formelen igjen, med $r = 2$ og $a = 2$:

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{\infty} xg(x)\mathrm{d}x = \int_0^\infty x \cdot 4xe^{-2x}\mathrm{d}x \\ &= 4\int_0^\infty x^2 e^{-2x}\mathrm{d}x = 4 \cdot \frac{2}{2^3} = 1.\end{aligned}$$

For å vise at det er 0.09 i sannsynlighet for mer enn to minutters forsinkelse, bruker vi delvis integrasjon.

$$\begin{aligned}\mathrm{P}(X > 2) &= \int_2^\infty 4xe^{-2x}\mathrm{dx} \\ &= 4 \cdot \left[-\frac{1}{2}xe^{-2x}\right]_{x=2}^{\infty} + 4 \cdot \int_2^\infty \frac{1}{2}e^{-2x}\mathrm{d}x \\ &= 4 \cdot \frac{1}{2} \cdot 2e^{-4} + \int_2^\infty 2e^{-2x}\mathrm{d}x \\ &= 4e^{-4} + e^{-4} = 5e^{-4} \approx 0.09.\end{aligned}$$

b) V er binomisk fordelt med $n = 22$ og $p = 0.09$ under forutsetning av at

- Hendelsene "mer enn 2 minutter forsinket" for to forskjellige dager er uavhengige.
- Sannsynligheten for "mer enn 2 minutter forsinket" er lik 0.09 hver dag.

Antall forsøk (dager) er bestemt på forhånd, det er to utfall og vi teller antall "suksesser". (Togselskapet ville neppe kalle en dag med mer enn to minutters forsinkelse for en suksess.)

$$\begin{aligned} \mathrm{P}(V \geq 2) &= 1 - \mathrm{P}(V \leq 1) = 1 - \mathrm{P}(V = 0) - \mathrm{P}(V = 1) \\ &= 1 - 0.91^{22} - 22 \cdot 0.91^{21} \cdot 0.09^{1} = 0.6012. \end{aligned}$$

Ser vi på et år med $n = 220$ virkedager, er $V \sim$ binomisk$(220, 0.09)$. Da kan vi bruke tilnærmingen til normalfordelingen, dvs

$$\begin{aligned} \mathrm{P}(V > 30) &= 1 - \mathrm{P}(V \leq 30) \approx 1 - \Phi\left(\frac{30 + 1/2 - 220 \cdot 0.09}{\sqrt{220 \cdot 0.09 \cdot (1 - 0.09)}}\right. \\ &= 1 - \Phi(2.52) = 0.0059. \end{aligned}$$

c) Setter $x = 2$ i den betingede fordelingen. Da har oppholdstiden Y fordeling $f(y|2) = e^{-y}$ for $y > 0$. Med andre ord er $Y|X = 2$ eksponensialfordelt med parameter (og forventningsverdi) 1.

Simultantetthet finner vi ved å multiplisere;

$$f(x, y) = f(y|x)g(x) = \frac{x}{2} e^{-\frac{xy}{2}} \cdot 4xe^{-2x} = 2x^2 e^{-x(2+\frac{y}{2})} \qquad \text{for } x > 0, y > 0.$$

Marginaltettheten for Y finnes ved å integrere ut x:

$$\begin{aligned} h(y) &= \int_0^\infty f(x, y)\mathrm{d}x = \int_0^\infty 2x^2 e^{-x(2+\frac{y}{2})}\mathrm{d}x \\ &= 2 \cdot \frac{2}{(2 + \frac{y}{2})^3} = \frac{32}{(4 + y)^3}, \qquad \text{for } y > 0. \end{aligned}$$

Her brukte vi enda en gang formelen som var oppgitt, denne gangen med $a = 2 + y/2$ og $r = 2$.

Oppgave 2

a) $Y \sim n(y; 500, 80)$. Transformerer Y til standard $N(0,1)$-normalfordeling.

$$\begin{aligned} \mathrm{P}(Y > 550) &= \mathrm{P}(\frac{Y-500}{80} > \frac{550-500}{80}) = \mathrm{P}(Z > \frac{5}{8}) \\ &= 1 - \mathrm{P}(Z \leq \frac{5}{8}) = 1 - \Phi(0.625) = 1 - 0.734 = 0.266. \end{aligned}$$

$Y_1 - Y_2 \sim n(y; 0, \sqrt{2} \cdot 80)$. (Lineærkombinasjonen av to uavhengige normalfordelinger er normalfordelt, sjekk forventningsverdi og varians ved de vanlige regnereglene.)

Da kan vi regne ut sannsynligheten for at målingene avviker med mer enn 80 g/tonn.

$$\begin{aligned} \mathrm{P}(|Y_1 - Y_2| > 80) &= 1 - \mathrm{P}(-80 < Y_1 - Y_2 < 80) \\ &= 1 - \mathrm{P}(\frac{-80}{80\sqrt{2}} < \frac{Y_1 - Y_2}{80\sqrt{2}} < \frac{80}{80\sqrt{2}}) \\ &= 1 - \mathrm{P}(-\frac{\sqrt{2}}{2} < Z < \frac{\sqrt{2}}{2}) = 2\mathrm{P}(Z \leq \frac{-\sqrt{2}}{2}) = 2\Phi(-0.707) \\ &= 2 \cdot 0.24 = 0.48. \end{aligned}$$

b) Setter inn $\overline{x} = 20$, $x_1 = \ldots = x_5 = 0$ og $x_6 = \ldots = x_{10} = 40$ i uttrykket for B.

$$\begin{aligned} B &= \frac{\sum_{j=1}^{5} -20Y_j + \sum_{j=6}^{10} 20Y_j}{\sum_{j=1}^{10} 20^2} = \frac{20\left(\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j\right)}{10 \cdot 20^2} \\ &= \frac{\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j}{200}, \text{som skulle vises.} \end{aligned}$$

$$A = \overline{Y} - B\overline{x} = \frac{1}{10}\sum_{j=1}^{10} Y_j - \frac{20}{200}\left(\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j\right) = \frac{1}{5}\sum_{j=1}^{5} Y_j.$$

A er skjæringspunktet regresjonslinja har med y-aksen. Det er kanskje ikke så rart at gjennomsnittet av målingene ved $x = 0$ er et estimat for denne verdien? (I hvert fall når målingene bare er gjort for to x-verdier.)

$$\mathrm{Var}(B) = \frac{1}{200^2}\left(\sum_{j=6}^{10} \mathrm{Var}(Y_j) + \sum_{j=1}^{5} \mathrm{Var}(Y_j)\right) = \frac{10\sigma^2}{200^2} = \frac{\sigma^2}{4000}.$$

c) Med bare to målepunkter, kan vi estimere variansen i hver ende for seg, dvs at vi beregner s_V^2 og s_E^2. (Husk at målingene ikke har samme forventningsverdi i de to endene av gruva, så vi kan ikke se på alle som ett datasett.) Ettersom vi antar samme varians i begge ender, er gjennomsnittet av s_V^2 og s_E^2 et godt estimat for σ^2.

Mer formelt, vi har en to-utvalgssituasjon, og kan da bruke s_p^2 fra pensum. Denne sikrer χ^2-fordeling og T-fordeling. Brukes estimatoren for variansen fra regresjonsanalysen, får en også samme resultat.

$$s^2 = \frac{1}{2}\left(s_V^2 + s_E^2\right) = \frac{1}{2}\left(\frac{\sum_{j=1}^{5}(y_j - \overline{y}_V)^2}{5-1} + \frac{\sum_{j=6}^{10}(y_j - \overline{y}_E)^2}{5-1}\right)$$
$$= \frac{1}{8}\left(\sum_{j=1}^{5}(y_j - \overline{y}_V)^2 + \sum_{j=6}^{10}(y_j - \overline{y}_E)^2\right) = \frac{26064 + 22720}{8} = 6098.$$

Hypotesene blir: H_0: $\beta = 12$ mot H_1: $\beta > 12$.

Vi baserer testen på estimatoren B. Siden variansen til B er ukjent, bruker vi estimatet $S_B^2 = \frac{s^2}{4000} = 1.525$ i stedet for $\frac{\sigma^2}{4000}$.

Testobservatoren, $\frac{B-12}{S_B}$, er T-fordelt med 8 frihetsgrader. Det er $n-2$ frihetsgrader denne gangen, fordi vi bruker "pooled" varians, eller, som sagt, variansestimatoren fra regresjonsanalysen. (Estimert varians er basert på to gjennomsnitt, $\overline{y}_V$ og $\overline{y}_E$. Da er det ikke så urimelig at vi mister to frihetsgrader?) Med oppgitte data blir stigningstallet

$$b = \frac{\sum_{j=6}^{10} y_j - \sum_{j=1}^{5} y_j}{200} = \frac{\overline{y}_E - \overline{y}_V}{40} = 17.$$

Gjennomfører hypotesetesten.

$$\frac{b-12}{s_B} = \frac{17-12}{\sqrt{1.525}} = 4.05 > t_{0.05,8} = 1.86,$$

som betyr at vi forkaster nullhypotesen på signifikansnivå 5%.

d) Fra det første uttrykket for B får vi

$$\mathrm{Var}(B) = \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

Variansen er liten for $\sum_{j=1}^{n}(x_j - \overline{x})^2$ stor. Altså vil vi ha alle $|x_j - \overline{x}|$ så store som mulig. Når $\overline{x}$ er fast, bør x_j-ene legges til endene, som i denne oppgaven. (Det kan være andre grunner til å spre målepunktene, f.eks. for å vurdere om dataene tilnærmet følger en rett linje, her var det antatt kjent.)

$$\mathrm{Var}(Y_0 - \widehat{Y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) = \frac{11}{10} \cdot \sigma^2$$

når $x_0 = \overline{x}$. Punktestimatet blir $\widehat{y}_0 = a + bx_0 = \overline{y}_V + 17 \cdot 20 = 470$.

Vi benytter fortsatt estimatet S^2 for σ^2, derfor fortsatt T-fordeling med $n - 2$ frihetsgrader. Prediksjonsintervallet blir derfor

$$(\widehat{y}_0 \pm t_{0.025,8} \cdot s\sqrt{\frac{11}{10}}) = (470 \pm 2.306 \cdot \sqrt{6098} \cdot \sqrt{1.1}) = (281.1, 658.9).$$

Den nye målingen, 600 g/tonn, ligger innenfor prediksjonsintervallet, så vi kan ikke konkludere med at den eller modellen er urimelig.

Oppgave 3

a) Estimatoren for avstanden a er gjennomsnittet av uavhengige identisk fordelte målinger, $n(x; a, \sigma_G)$.

$$\mathrm{E}(\widehat{a}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} a = a.$$
$$\mathrm{Var}(\widehat{a}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_G^2 = \frac{\sigma_G^2}{n}.$$

Estimatoren er en lineærkombinasjon av Gaussiske (Normalfordelte) tilfeldige variable, og er derfor Gaussisk; $n(\widehat{a}; a, \frac{\sigma_G}{\sqrt{n}})$.

b) Siden σ er ukjent, er $\frac{\widehat{a}-a}{S_G/\sqrt{n}}$ T-fordelt med $n-1$ frihetsgrader. Da skal vi ha

$$\mathrm{P}(-t_{n-1,0.025} < T < t_{n-1,0.025}) = 0.95$$
$$\mathrm{P}(-t_{n-1,0.025} < \frac{\widehat{a}-a}{S_G/\sqrt{n}} < t_{n-1,0.025}) = 0.95$$
$$\mathrm{P}(\widehat{a} - t_{n-1,0.025}\frac{S_G}{\sqrt{n}} < a < \widehat{a} + t_{n-1,0.025}\frac{S_G}{\sqrt{n}}) = 0.95.$$

Med andre ord er 95%-intervallestimatoren for a $\left[\widehat{a} - t_{n-1,0.025}\frac{S_G}{\sqrt{n}}, \widehat{a} + t_{n-1,0.025}\frac{S_G}{\sqrt{n}}\right]$.

c) For å finne SME for a og σ_M, begynner vi med Likelihood-funksjonen. Setter $\sigma = \sigma_M = \sigma_G/4$.

$$\begin{aligned} L(x_1,\ldots,x_n,y_1,\ldots,y_m;a,\sigma) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_G}e^{-\frac{(x_i-a)^2}{2\sigma_G^2}}\cdot\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma_M}e^{-\frac{(y_i-a)^2}{2\sigma_M^2}} \\ &= \frac{1}{\sqrt{2\pi}^n\sigma_G^n}e^{-\frac{\sum_{i=1}^{n}(x_i-a)^2}{2\sigma_G^2}}\frac{1}{\sqrt{2\pi}^m\sigma_M^m}e^{-\frac{\sum_{i=1}^{m}(y_i-a)^2}{2\sigma_M^2}} \\ &= \frac{1}{(2\pi)^{(m+n)/2}(4\sigma)^n\sigma^m}e^{-\frac{\sum_{i=1}^{n}(x_i-a)^2}{2(4\sigma)^2}-\frac{\sum_{i=1}^{m}(y_i-a)^2}{2\sigma^2}}. \end{aligned}$$

Tar logaritmen for å forenkle deriveringen som kommer;

$$\begin{aligned} l(x_1,\ldots,x_n,&y_1,\ldots,y_m;a,\sigma) = \ln(L(\ldots)) \\ &= -\frac{n+m}{2}\ln(2\pi) - n\ln(4\sigma) - m\ln(\sigma) - \frac{\sum_{i=1}^{n}(x_i-a)^2}{32\sigma^2} - \frac{\sum_{i=1}^{m}(y_i-a)^2}{2\sigma^2}. \end{aligned}$$

SME for a og σ_M krever $\frac{\partial l}{\partial a} = 0 = \frac{\partial l}{\partial \sigma}$. Begynner med å derivere mhp a.

$$\begin{aligned} \frac{\partial l}{\partial a} &= \frac{2}{32\sigma^2}\sum_{i=1}^{n}(x_i-a) + \frac{2}{2\sigma^2}\sum_{i=1}^{m}(y_i-a) \\ &= \frac{1}{\sigma^2}\left[\frac{1}{16}\sum_{i=1}^{n}(x_i-a) + \sum_{i=1}^{m}(y_i-a)\right]. \end{aligned}$$

Setter inn a_{SME} for a og uttrykket over lik null. Det gir

$$\begin{aligned} \frac{1}{16}\sum_{i=1}^{n}(x_i-a_{\text{SME}}) + \sum_{i=1}^{m}(y_i-a_{\text{SME}}) &= 0 \\ \frac{1}{16}\sum_{i=1}^{n}x_i - \frac{n}{16}a_{\text{SME}} + \sum_{i=1}^{m}y_i - ma_{\text{SME}} &= 0 \\ \left(\frac{n}{16}+m\right)a_{\text{SME}} = \frac{1}{16}\sum_{i=1}^{n}x_i + \sum_{i=1}^{m}y_i &= \frac{n\overline{x}}{16} + m\overline{y} \\ a_{\text{SME}} &= \frac{n\overline{X}+16m\overline{Y}}{n+16m}. \end{aligned}$$

Så gjenstår derivasjon mhp σ.

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{m}{\sigma} + \frac{2\sum_{i=1}^{n}(x_i-a)^2}{32\sigma^3} + \frac{2\sum_{i=1}^{m}(y_i-a)^2}{2\sigma^3}.$$

Igjen setter vi dette lik null for $a = a_{\text{SME}}$ og $\sigma = \sigma_{\text{SME}}$:

$$\begin{aligned} -\frac{n}{\sigma_{\text{SME}}} - \frac{m}{\sigma_{\text{SME}}} + \frac{\sum_{i=1}^{n}(x_i-a_{\text{SME}})^2}{16\sigma_{\text{SME}}^3} + \frac{\sum_{i=1}^{m}(y_i-a_{\text{SME}})^2}{\sigma_{\text{SME}}^3} &= 0 \\ (n+m)\sigma_{\text{SME}}^2 = \frac{1}{16}\sum_{i=1}^{n}(x_i-a_{\text{SME}})^2 + \sum_{i=1}^{m}(y_i-a_{\text{SME}})^2 & \\ \sigma_{\text{SME}}^2 = \frac{1}{n+m}\left[\frac{1}{16}\sum_{i=1}^{n}(X_i-a_{\text{SME}})^2 + \sum_{i=1}^{m}(Y_i-a_{\text{SME}})^2\right]. & \end{aligned}$$

Dette siste uttrykket blir på ingen måte enklere om en i tillegg setter inn for a_{SME}.