Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

LØSNINGSFORSLAG TILEKSAMEN I FAG TMA4240/TMA4245 STATISTIKK

10. august 2005

Oppgave 1 Smeltepunktsbestemmelse

a) Vi jobber i dette punktet med uavhengige og identisk fordelte stokastiske variabler fra en normalfordeling med forventningsverdi $\mu = 1468°$C og standardavvik $\sigma = 2°$ C.

La X være en slik normalfordelt stokastisk variabel, og la Z være en standard normalfordelt stokastisk variabel.

$$\begin{aligned} P(X < 1467) &= P(X \leq 1467) = P\left(\frac{X-1468}{2} \leq \frac{1467-1468}{2}\right) \\ &= P(Z \leq -\frac{1}{2}) = \Phi(-\frac{1}{2}) = \underline{\underline{0.3085}} \end{aligned}$$

Anta så at metallurgen tar åtte uavhengige målinger av smeltepunktet, $X_1, X_2, ..., X_8$.

Da er gjennomsnittet, $\bar{X} = \frac{1}{8}\sum_{i=1}^{8} X_i$, også en normalfordelt stokastisk variabel (siden det er en lineærkombinasjon av uavhengige og normalfordelte stokastiske variabler), med forventning og varians:

$$\begin{aligned} \mathrm{E}(\bar{X}) &= \mathrm{E}(\frac{1}{8}\sum_{i=1}^{8} X_i) = \frac{1}{8}\sum_{i=1}^{8}\mathrm{E}(X_i) = \frac{1}{8}\sum_{i=1}^{8}\mu = \mu = 1468 \\ \mathrm{Var}(\bar{X}) &= \mathrm{Var}(\frac{1}{8}\sum_{i=1}^{8} X_i) = \frac{1}{8^2}\sum_{i=1}^{8}\mathrm{Var}(X_i) = \frac{1}{8^2}\sum_{i=1}^{8}\sigma^2 = \frac{\sigma^2}{8} = \frac{2^2}{8} = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} P(1467 < \bar{X} < 1469) &= P(\bar{X} \leq 1469) - P(\bar{X} \leq 1467) \\ &= P\left(\frac{\bar{X}-1468}{\sqrt{\frac{1}{2}}} \leq \frac{1469-1468}{\sqrt{\frac{1}{2}}}\right) - P\left(\frac{\bar{X}-1468}{\sqrt{\frac{1}{2}}} \leq \frac{1467-1468}{\sqrt{\frac{1}{2}}}\right) \\ &= P(Z \leq \sqrt{2}) - P(Z \leq -\sqrt{2}) = \Phi(1.41) - \Phi(-1.41) \\ &= 0.9207 - 0.0793 = \underline{\underline{0.84}} \end{aligned}$$

Vi skal i resten av oppgaven anta at forventningsverdien til legeringens smeltepunkt, μ, er ukjent, men at standardavvik er kjent og lik $\sigma = 2°$ C.

b) Utled et 90% konfidensinterval for μ basert på n uavhengige målinger av smeltepunktet, $X_1, X_2, ..., X_n$.

Her er σ kjent og vi baserer oss på at

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

er standard normalfordelt.

Her lar vi $\alpha = 0.1$ for å få et $(1-\alpha)100\% = 90\%$ konfidensintervall.

$$\begin{aligned} P(-z_{\frac{\alpha}{2}} &\leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha \\ P(-z_{\frac{\alpha}{2}} &\leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}) = 1-\alpha \\ P(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha \\ P(\hat{\mu}_L &\leq \mu \leq \hat{\mu}_U) = 1-\alpha \end{aligned}$$

Lengden til intervallet er gitt som

$$\begin{aligned} L &= \hat{\mu}_U - \hat{\mu}_L = (\bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) - (\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) \\ &= 2 \cdot z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

Vi observerer at lengden av intervallet ikke er stokastisk når σ er kjent.

Vi skal finne minste n slik at lengden på intervallet ikke overstiger $3°$C.

$$\begin{aligned} L &\leq 3 \\ 2 \cdot z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &\leq 3 \\ n &\geq (2 \cdot z_{\frac{\alpha}{2}}\frac{\sigma}{3})^2 \end{aligned}$$

Innsatt $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$ og $\sigma = 2$ får vi

$$n \geq (2 \cdot 1.645 \cdot \frac{2}{3})^2 = 4.8$$

Det minste antall observasjoner som gir et intervall med lengde som ikke overskrider $3°$C er $n_0 = 5$.

Vi bruker de $n_0 = 5$ første observasjonene gitt i tabellen i oppgaveteksten til å bestemme intervallet numerisk. Finner at $\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 1468.88$.

$$\hat{\mu}_L = \bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n_0}} = 1467.41$$
$$\hat{\mu}_U = \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n_0}} = 1470.35$$

Som kvalitetssjekk observerer vi at lengden på intervallet da er $1470.35 - 1467.41 = 2.94$ som er mindre enn 3.

c) Vi lar metallurgens lange erfaring med legeringens smeltepunkt være den konservative hypotesen, nullhypotesen, som Professor Stål forsøker å bestride. Professor Ståls angrep blir da den alternative hypotesen.

Vi kaller $\mu_0 = 1468°$C, og får følgende hypoteser:

$$H_0 : \mu = \mu_0 \qquad vs. \qquad H_1 : \mu > \mu_0$$

Vi bruker $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ som estimator for μ og vi vil forkaste H_0 når $\bar{X}$ er stor (fordi da tror vi på den alternative hypotesen). Vi vet at under H_0 så er

$$Z_0 = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

standard normalfordelt. Hvis vi skal forkaste H_0 når $\bar{X}$ er stor, vil vi også forkaste H_0 når Z_0 er stor og vi bestemmer oss for å forkaste H_0 når $Z_0 > k$ (Z_0 er dermed testobservatoren vår). Videre bestemmer vi k slik at

$$P(\text{type I feil}) = P(\text{forkaste } H_0|H_0 \text{ er sann}) \leq \alpha.$$

Innsatt $Z_0 > k$ for hendelsen "forkaste H_0" og μ_0 for hendelsen "H_0 sann":

$$P(\text{forkaste } H_0|H_0 \text{ er sann}) \leq \alpha$$
$$P(Z_0 > k|\mu_0) \leq \alpha$$
$$P(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > k|\mu_0) \leq \alpha$$

Tallet k som har areal α til høyre i standard normalfordelingen er kvantilen z_α, dvs. $k = z_\alpha$. Dvs. vi forkaster H_0 når

$$\underline{\underline{Z_0 = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha}}$$

For $\alpha = 0.05$ er $z_{0.05} = 1.645$. Videre har vi $z_0 = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{1469.0-1468}{\frac{2}{\sqrt{8}}} = 1.41$, som er mindre enn 1.645 og dermed gir det *ikke* forkastning. Konklusjonen er at det vi har observert (eller noe verre) er ganske sannsynlig (har høyere sannsynlighet enn 0.05) når H_0 er sann, og vi forkaster dermed ikke H_0.

Nå vil vi se hvor lett det er å forkaste H_0 med regelen vår hvis i virkeligheten $\mu = 1470$. Denne sannsynligheten er avhengig av vårt valgte signifikansnivå og hvor mange observasjoner vi har brukt til å lage testen vår (vi har et større forkastningsområde når vi har mange observasjoner). Vi skal finne et minimum antall observasjoner, n, slik at
$P(\text{Forkaste } H_0|H_0 \text{ er gal og i virkeligheten } \mu = 1470) \geq 0.99$. Utfordringen vår er nå at Z_0 ikke lenger er standard normalfordelt (siden μ_0 ikke er den riktige forventningen), slik at vi må bruke at $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ innsatt $\mu = 1470$ er standard normafordelt.

$$P(\text{Forkaste } H_0|H_0 \text{ gal og i virkeligheten } \mu = 1470) \geq 0.99$$
$$P(Z_0 > z_\alpha|\mu = 1470) \geq 0.99$$
$$P(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha|\mu = 1470) \geq 0.99$$
$$P(\bar{X} > \mu_0 + z_\alpha\frac{\sigma}{\sqrt{n}}|\mu = 1470) \geq 0.99$$
$$P(\frac{\bar{X} - 1470}{\frac{\sigma}{\sqrt{n}}} > \frac{\mu_0 - 1470}{\frac{\sigma}{\sqrt{n}}} + z_\alpha|\mu = 1470) \geq 0.99$$

Når $\mu = 1470$ er $\frac{\bar{X}-1470}{\frac{\sigma}{\sqrt{n}}}$ i ligningen over standard normalfordelt, og vi kan matche $\frac{\mu_0-1470}{\frac{\sigma}{\sqrt{n}}} + z_\alpha$ med kvantilen i standard normalfordelingen som har areal 0.99 til høyre for seg (siden det står $>$ i ulikheten). Areal 0.99 til høyre har $-z_{0.01}$ kvantilen. Vi setter inn for $\mu_0 = 1468$, $\sigma = 2$ og $\alpha = 0.05$, slik at vi fokuserer på at det er kun n som er ukjent. I første overgangen har vi delt på et negativt tall og må snu ulikhetstegnet.

$$\frac{1468 - 1470}{\frac{2}{\sqrt{n}}} + z_{0.05} \leq -z_{0.01}$$
$$\sqrt{n} \geq \frac{-(z_{0.01} + z_{0.05}) \cdot 2}{1468 - 1470}$$
$$n \geq (\frac{-(z_{0.01} + z_{0.05}) \cdot 2}{1468 - 1470})^2$$
$$n \geq (\frac{-(2.326 + 1.645) \cdot 2}{1468 - 1470})^2 = 15.77$$

Det minste antallet observasjoner som gir styrke 0.99 i alternativet $\mu = 1470$ er $n = 16$ observasjoner.

Oppgave 2 Ventetid på snekkertjenester

a) Sannsynligheten for at ventetiden er lenger enn 2 uker:

$$P(X>2)=1-P(X\le 2)=1-F(2)=1-(1-\exp(-0.04\cdot 2^2))=\exp(-0.16)=\underline{\underline{0.8521}}.$$

Sannsynligheten for at du må vente minst 5 uker, gitt at du må vente i minst 2 uker:

$$\begin{aligned}P(X>5|X>2) &= \frac{P(X>5\cap X>2)}{P(X>2)}=\frac{P(X>5)}{P(X>2)}=\frac{1-P(X\le 5)}{P(X>2)}\\ &= \frac{1-F(5)}{P(X>2)}=\frac{1-(1-\exp(-0.04\cdot 5^2))}{0.8521}=\frac{0.3679}{0.8521}=\underline{\underline{0.4317}}.\end{aligned}$$

Sannsynlighetstettheten til X for $x\ge 0$ finner vi ved å derivere $F(x)$:

$$f(x)=\frac{dF(x)}{dx}=0-(-2\alpha x\exp(-\alpha x^2))=2\alpha x\exp(-\alpha x^2),\quad \text{for}\quad x\ge 0.$$

b) SME for α:

Finner først rimelighetsfunksjonen, som er simultanfordelingen for $X_1,X_2,\ldots,X_n$ sett på som funksjon av x_i'ene og α:

$$\begin{aligned}L(x_1,x_2,\ldots,x_n;\alpha) &= f(x_1,x_2,\ldots,x_n;\alpha)=\prod_{i=1}^n f(x_i;\alpha)\\ &= \prod_{i=1}^n 2\alpha x_i\exp(-\alpha x_i^2)=2^n\alpha^n(\prod_{i=1}^n x_i)\ \exp(-\alpha\sum_{i=1}^n x_i^2).\end{aligned}$$

Tar logaritmen:

$$l(x_1,x_2,\ldots,x_n;\alpha)=\ln L(x_1,x_2,\ldots,x_n;\alpha)=n\ln 2+n\ln\alpha+\sum_{i=1}^n \ln x_i-\alpha\sum_{i=1}^n x_i^2.$$

Deriverer med hensyn på α og setter lik 0:

$$\begin{aligned}\frac{\partial l(x_1,x_2,\ldots,x_n;\alpha)}{\partial\alpha}=0+\frac{n}{\alpha}+0-\sum_{i=1}^n x_i^2 &= 0\\ \frac{n}{\alpha} &= \sum_{i=1}^n x_i^2\\ \alpha &= \frac{n}{\sum_{i=1}^n x_i^2}.\end{aligned}$$

Dvs. SME for α er $\alpha^*=\frac{n}{\sum_{i=1}^n x_i^2}$, som er ulik $\hat\alpha$. $\hat\alpha$ er dermed ikke SME for α.

Estimator $\hat\mu$ for μ: $\hat\mu=\frac{\sqrt{\pi}}{2\sqrt{\hat\alpha}}=\frac{\sqrt{\pi}\sqrt{\sum_{i=1}^n X_i^2}}{2(n-1)}$.

Innsatt verdier blir $\hat\alpha=0.029$ og estimatet for μ blir $\hat\mu=\frac{\sqrt{\pi}}{2\sqrt{\hat\alpha}}=\frac{\sqrt{\pi}}{2\sqrt{0.029}}=5.2$ uker.

c) Sannsynlighetsfordelingen for $Y=X^2$ finner vi ved å bruke transformasjonsformelen.
La $Y=u(X)=X^2$ slik at $X=w(Y)=\sqrt{Y}$ (har at $X>0$). Dermed er

$$f_Y(y)=f_X(w(y))\,|w'(y)|=2\alpha\sqrt{y}\exp(-\alpha(\sqrt{y})^2)\,|\frac{1}{2\sqrt{y}}|=\alpha\exp(-\alpha y).$$

Dette er sannsynlighetstettheten i eksponensialfordelingen med forventning $1/\alpha$, og dermed har vi vist at Y er eksponensialfordelt.

Forventningsverdi for $\hat\alpha$:

$$\mathrm{E}(\hat\alpha)=\mathrm{E}\left(\frac{n-1}{\sum_{i=1}^n X_i^2}\right)=(n-1)\cdot\mathrm{E}\left(\frac{1}{\sum_{i=1}^n X_i^2}\right)=(n-1)\cdot\frac{1}{(1/\alpha)(n-1)}=\frac{n-1}{n-1}\,\alpha=\alpha.$$

Her har vi brukt resultatet oppgitt i oppgaveteksten.

Siden $\mathrm{E}(\hat\alpha)=\alpha$, er $\hat\alpha$ forventningsrett.

Oppgave 3 **Ulykker**

a) Vi har at X_i er Poisson-fordelt med $E(X_i)=a_i\lambda$. I Poisson-fordelingen er også $\mathrm{Var}(X_i)=a_i\lambda$.

Forventningsverdien til de to estimatorene $\hat\Lambda$ og Λ^*

$$\begin{aligned}\mathrm{E}(\hat\Lambda) &= \frac{1}{n}\sum_{i=1}^n\frac{1}{a_i}\mathrm{E}(X_i)=\frac{1}{n}\sum_{i=1}^n\frac{1}{a_i}a_i\lambda=\frac{1}{n}\cdot n\cdot\lambda=\underline{\underline{\lambda}}\\ \mathrm{E}(\Lambda^*) &= \frac{\sum_{i=1}^n\mathrm{E}(X_i)}{\sum_{i=1}^n a_i}=\frac{\sum_{i=1}^n a_i\lambda}{\sum_{i=1}^n a_i}=\lambda\cdot\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n a_i}=\underline{\underline{\lambda}}\end{aligned}$$

Variansen til de to estimatorene $\hat{\Lambda}$ og Λ^*

$$\begin{aligned}
\mathrm{Var}(\hat{\Lambda}) &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{a_i^2}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{a_i^2}a_i\lambda = \underline{\underline{\lambda\cdot\frac{1}{n^2}\cdot\sum_{i=1}^{n}\frac{1}{a_i}}} \\
\mathrm{Var}(\Lambda^*) &= \frac{\sum_{i=1}^{n}\mathrm{Var}(X_i)}{(\sum_{i=1}^{n}a_i)^2} = \frac{\sum_{i=1}^{n}a_i\lambda}{(\sum_{i=1}^{n}a_i)^2} = \underline{\underline{\lambda\cdot\frac{1}{\sum_{i=1}^{n}a_i}}}
\end{aligned}$$

Numeriske verdier for $\hat{\Lambda}$ og Λ^* fra data i tabell 2.

$$\begin{aligned}
\hat{\lambda} &= \frac{1}{5}\sum_{i=1}^{5}\frac{x_i}{a_i} = \frac{1}{5}\cdot 4.33\cdot 10^{-5} = \underline{\underline{8.67\cdot 10^{-6}}} \\
\lambda^* &= \frac{\sum_{i=1}^{5}x_i}{\sum_{i=1}^{5}a_i} = \frac{8}{9.0\cdot 10^5} = \underline{\underline{8.89\cdot 10^{-6}}}
\end{aligned}$$

Hvilken av de to estimatorene ville du foretrekke i denne situasjonen? Begge estimatorene er forventningsrette, men de har ulik varians. Numerisk varians (på konstant nær pga. at λ er ukjent):

$$\begin{aligned}
\mathrm{Var}(\hat{\Lambda}) &= \lambda\cdot\frac{1}{n^2}\cdot\sum_{i=1}^{n}\frac{1}{a_i} = 1.63\cdot 10^{-6}\cdot\lambda \\
\mathrm{Var}(\Lambda^*) &= \lambda\cdot\frac{1}{\sum_{i=1}^{n}a_i} = 1.11\cdot 10^{-6}\cdot\lambda
\end{aligned}$$

Begge estimatorene er forventningsrette, men estimatoren Λ^* har minst varians. Derfor så foretrekker vi Λ^*.

b) Fordelingen til Y er den betingede fordelingen til X gitt at $X \geq 1$ (minst én stor ulykke). Vi benytter i utledningen at X er Poisson-fordelt med punktsannsynlighet

$$P(X = x) = \frac{(a\lambda)^x e^{-a\lambda}}{x!}.$$

Dermed har vi:

$$\begin{aligned}
P(Y = y) &= P(X = y|X \geq 1) = \frac{P(X = y \cap X \geq 1)}{P(X \geq 1)} = \frac{P(X = y)}{P(X \geq 1)} = \frac{P(X = y)}{1 - P(X = 0)} \\
&= \frac{\frac{(a\lambda)^y}{y!}e^{-a\lambda}}{1 - \frac{(a\lambda)^0}{0!}e^{-a\lambda}} = \frac{(a\lambda)^y}{y!}\cdot\frac{e^{-a\lambda}}{1 - e^{-a\lambda}}
\end{aligned}$$

som skulle vises.

Forventningsverdien til Y finner vi fra $\mathrm{E}(Y) = \sum_{y=1}^{\infty} y \cdot P(Y = y)$, innsatt $P(Y = y)$.

$$\begin{aligned}
\mathrm{E}(Y) &= \sum_{y=1}^{\infty} y\cdot P(Y = y) = \sum_{y=1}^{\infty} y\cdot\frac{(a\lambda)^y}{y!}\cdot\frac{e^{-a\lambda}}{1 - e^{-a\lambda}} \\
&= \sum_{y=1}^{\infty}\frac{(a\lambda)^y}{(y-1)!}\cdot\frac{e^{-a\lambda}}{1 - e^{-a\lambda}} = \sum_{z=0}^{\infty}\frac{(a\lambda)^{z+1}}{z!}\cdot\frac{e^{-a\lambda}}{1 - e^{-a\lambda}} \\
&= \frac{a\lambda}{1 - e^{-a\lambda}}\sum_{z=0}^{\infty}\frac{(a\lambda)^z}{z!}e^{-a\lambda} = \frac{a\lambda}{1 - e^{-a\lambda}}\cdot 1 = \frac{a\lambda}{1 - e^{-a\lambda}}
\end{aligned}$$

Skiftet av sum fra $y = (1, \infty)$ til $z = y - 1$ der $z = (0, \infty)$ er for å få $\sum_{z=0}^{\infty}\frac{(a\lambda)^z}{z!}e^{-a\lambda} = 1$ siden dette er sum over alle punktsannsynligheter i en Poisson-fordeling med forventning $a\lambda$.