**Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag**

LØSNINGSFORSLAG TIL EKSAMEN I FAG TMA4245 STATISTIKK
Fredag 19.mai 2006

Oppgave 1 Feil på mobilnett

a) Den kumulative fordelingsfunksjonen $F(x) = P(X \le x)$ beregner vi ved å integrere sannsynlighetstettheten $f(x)$. Dvs.

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{1}^{x} \beta t^{-\beta-1}dt = \beta\,\frac{1}{-\beta}\left[t^{-\beta}\right]_1^x = (-1)\left[x^{-\beta}-1\right] = 1 - x^{-\beta}.$$

Sannsynligheten for at det går mer enn 2 uker mellom to påfølgende feil, når $\beta = 3$, er

$$P(X > 2) = 1 - P(X \le 2) = 1 - F(2) = 1 - (1 - 2^{-\beta}) = 2^{-3} = 0.125$$

Sannsynligheten for at nettet svikter før det er gått 3.5 uker, gitt at det har gått minst 2 uker siden siste feil, er (med $\beta = 3$)

$$\begin{aligned} P(X \le 3.5 \mid X > 2) &= \frac{P(X \le 3.5 \cap X > 2)}{P(X > 2)} = \frac{P(X \le 3.5) - P(X \le 2))}{P(X > 2)} \\ &= \frac{F(3.5) - F(2)}{1 - F(2)} = \frac{(1 - 3.5^{-3}) - (1 - 2^{-3})}{0.125} = 0.813 \end{aligned}$$

b) Sannsynlighetsmaksimeringsestimatoren, SME for β:

Simultantettheten for $X_1, \ldots, X_n$ er $f(x_1, \ldots, x_n; \beta) \overset{\text{uavh.}}{=} \prod_{i=1}^n f(x_i; \beta) = \prod_{i=1}^n \beta x_i^{-\beta-1}$. Rimelighetsfunksjonen er simultanfordelingen sett på som funksjon av β, og kan skrives som

$$L(x_1, \ldots, x_n; \beta) = \beta^n \prod_{i=1}^n x_i^{-\beta-1}.$$

SME er den verdien for β som maksimerer $L(x_1, \ldots, x_n; \beta)$. Denne verdien finner vi ved først å ta logaritmen, så derivere og sette lik 0:

$$\begin{aligned} l(x_1, \ldots, x_n; \beta) &= \ln\left(L(x_1, \ldots, x_n; \beta)\right) = \ln\left(\beta^n \prod_{i=1}^n x_i^{-\beta-1}\right) \\ &= \ln(\beta^n) + \ln\left(\prod_{i=1}^n x_i^{-\beta-1}\right) \\ &= n\ln(\beta) + \sum_{i=1}^n(-(\beta+1)\ln(x_i)) = n\ln(\beta) - (\beta+1)\sum_{i=1}^n \ln(x_i) \\ \frac{dl(x_1, \ldots, x_n; \beta)}{d\beta} &= \frac{n}{\beta} - \sum_{i=1}^n \ln(x_i) = 0 \\ \beta &= \frac{n}{\sum_{i=1}^n \ln(x_i)} \end{aligned}$$

Dette gir at SME for β er

$$\hat\beta = \hat\beta_1 = \frac{n}{\sum_{i=1}^n \ln X_i}.$$

Når vi setter inn de observerte verdiene får vi følgende estimat for β:

$$\hat\beta = \hat\beta_1 = \frac{n}{\sum_{i=1}^n \ln x_i} = \frac{10}{3.39} = 2.95.$$

c) Vi skal først vise at $2\beta\ln(X_i)$ er kjikvadratfordelt med 2 frihetsgrader (som er det samme som en eksponensialfordeling).
La $Y_i = 2\beta\ln(X_i)$. Vi kan finne sannsynlighetsfordelingen til Y_i ved å bruke transformasjonsformelen (vi ser her bort fra indeksen i i utledningen). La

$$\begin{aligned} y &= u(x) = 2\beta\ln(x), \quad \text{slik at} \\ x &= w(y) = \exp(\frac{y}{2\beta}). \end{aligned}$$

La $f_Y(y)$ være sannsynlighetstettheten til Y. Transformasjonsformelen sier da at

$$f_Y(y) = f_X(w(y))|w'(y)|.$$

Vi deriverer $w(y)$ og får $w'(y) = \frac{1}{2\beta}\exp(\frac{y}{2\beta})$. Sannsynlighetstettheten til Y blir da

$$\begin{aligned} f_Y(y) &= f_X(w(y))|w'(y)| = \beta(\exp(\frac{y}{2\beta}))^{-\beta-1}\frac{1}{2\beta}\exp(\frac{y}{2\beta}) \\ &= \frac{1}{2}\exp((-\beta-1)\frac{y}{2\beta})\exp(\frac{y}{2\beta}) = \frac{1}{2}\exp(-\beta\frac{y}{2\beta} - \frac{y}{2\beta} + \frac{y}{2\beta}) \\ &= \frac{1}{2}\exp(-\frac{y}{2}). \end{aligned}$$

Uttrykket for $f_Y(y)$ kan skrives

$$f_Y(y) = \frac{1}{2^{2/2}\Gamma(2/2)} y^{2/2-1} \exp(-\frac{y}{2}),$$

siden $\Gamma(2/2) = \Gamma(1) = 1$. Dette er sannsynlighetstettheten for en kjikvadratfordelt stokastisk variabel med 2 frihetsgrader. Dermed har vi vist at $Y_i = 2\beta \ln(X_i)$ er kjikvadratfordelt med 2 frihetsgrader, dvs. $Y_i \sim \chi^2_2$.

La $Z = 2\beta \sum_{i=1}^n \ln(X_i)$. Med $Y_i = 2\beta \ln(X_i)$ har vi at

$$Z = 2\beta \sum_{i=1}^n \ln(X_i) = \sum_{i=1}^n 2\beta \ln(X_i) = \sum_{i=1}^n Y_i.$$

Vi har vist at $Y_i \sim \chi^2_2$, og siden en sum av uavhengige kjikvadratfordelte stokastiske variabler er kjikvadratfordelt, med summen av frihetsgradene, er $Z = 2\beta \sum_{i=1}^n \ln(X_i)$ kjikvadratfordelt med $\sum_{i=1}^n 2 = 2n$ frihetsgrader.

Konfidensintervall for β:

Vi bruker at $Z = 2\beta \sum_{i=1}^n \ln(X_i) \sim \chi^2_{2n}$. La $\alpha = 0.05$. Vi får da at

$$\begin{aligned}
P(\chi^2_{1-\alpha/2,2n} < Z < \chi^2_{\alpha/2,2n}) &= 1-\alpha \\
P(\chi^2_{1-\alpha/2,2n} < 2\beta \sum_{i=1}^n \ln(X_i) < \chi^2_{\alpha/2,2n}) &= 1-\alpha \\
P(\frac{\chi^2_{1-\alpha/2,2n}}{2\sum_{i=1}^n \ln(X_i)} < \beta < \frac{\chi^2_{\alpha/2,2n}}{2\sum_{i=1}^n \ln(X_i)}) &= 1-\alpha
\end{aligned}$$

Et 95% konfidensintervall for β blir da

$$\left[\frac{\chi^2_{1-0.025,2n}}{2\sum_{i=1}^n \ln(x_i)} < \beta < \frac{\chi^2_{0.025,2n}}{2\sum_{i=1}^n \ln(x_i)}\right]$$

Insatt observerte verdier får vi

$$\left[\frac{9.591}{2\cdot 3.39} < \beta < \frac{34.170}{2\cdot 3.39}\right] = [1.41, 5.04].$$

Oppgave 2 Transport av masse

X er hypoergeometrisk fordelt med $N = 1000$ turer, $k = 5$ turer kjører transportfirmaet gjennom sentrum og $N - k = 995$ utenom sentrum, og vi tar en stikkprøve av størrelse $n = 5$.

Betingelser:

- Et tilfeldig utvalg av størrelse n tas *uten* tilbakelegging fra N enheter. Her: et tilfeldig utvalg av $n = 5$ turer sjekket blant N turer som totalt kjøres.
- De N enhetene deles inn i to grupper, k suksesser og $N - k$ fiaskoer. Her: $k = 5$ turer kjøres gjennom sentrum og $N - k = 995$ turer kjøres utenom sentrum.
- X er antallet suksesser blant de n. Her: X er antall turer gjennom sentrum av de $n = 5$ turene som ble sjekket.

Punktsannsynligheten i hypergeometrisk fordeling, $N = 1000, k = 5, n = 5$ er gitt som:

$$P(X = x) = \frac{\binom{5}{x}\binom{995}{5-x}}{\binom{1000}{5}}$$

og mulige verdier for x er $0, 1, 2, 3, 4, 5$.

$$P(X = 0) = \frac{\binom{5}{0}\binom{995}{5-0}}{\binom{1000}{5}} = 0.9752$$

Siden $P(X = 0) = 0.975$ må $x = 0$ være den verdien av x som gir høyest punktsannsynlighet (siden summen av alle punktsannsynligheter er 1 kan ingen annen punktsannsynlighet være større enn 1-0.975).

$$P(X = 5) = \frac{\binom{5}{5}\binom{995}{0}}{\binom{1000}{5}} = \underline{\underline{1.21 \cdot 10^{-13}}}$$

Til sammenligning er sannsynligheten for å vinne 7 rette i lotto $1.85 \cdot 10^{-7}$.

Kommentar 1: Når N er stor i forhold til n (boka nevner som tommelfingerregel at $n/N \leq 0.05$, og her er jo $5/1000 = 0.005$) så kan binomisk fordeling brukes som en tilnærming til hypergeometrisk fordeling når vi regner ut sannsynligheter. Da gjør vi $n = 5$ forsøk og i hvert forsøk sjekker vi om transporten skjer gjennom bykjernen, $p = \frac{k}{N} = \frac{5}{1000}$ er sannsynlighet for transport gjennom bykjernen, og X er antall transporter gjennom bykjernen for de $n = 5$ undersøkt. Da kan punktsansynligheten til X tilnærmes med

$$P(X = x) = \binom{n}{x}(\frac{k}{N})^x(1 - \frac{k}{N})^{n-x} = \binom{5}{x}(\frac{5}{1000})^x(1 - \frac{5}{1000})^{5-x}$$

Videre er tilnærmet:

$$\begin{aligned} P(X=0) &= \binom{5}{0}(\frac{5}{1000})^0(1-\frac{5}{1000})^{5-0} = (1-\frac{5}{1000})^5 = 0.975 \\ P(X=5) &= \binom{5}{5}(\frac{5}{1000})^5(1-\frac{5}{1000})^{5-5} = (\frac{5}{1000})^5 = 3.125\cdot 10^{-12} \end{aligned}$$

Kommentar 2: Denne oppgaven er basert på en henvendelse fra en tidligere bygg-student, og er basert på faktiske forhold. Dog, transportfirmaet sa først at alle 1000 turene var kjørt utenom bykjernen og kun etter at de be møtt med fakta på at stikkprøve av 5 turer viste transport gjennom bykjernen så informerte de om at det kun var akkurat disse 5 turene (av de 1000) som hadde blitt kjørt gjennom bykjernen. La oss tenke oss at vi ser på dette som en hypotesetest, der vi ønsker å finne ut om det er grunn til å tro at transportfirmaet har kjørt mer enn $k=5$ av turene gjennom bykjernen:

$$H_0: k=5 \text{ vs. } H_1: k>5$$

P-verdien til testen ville vært å regne ut $P(X=5)$ som vi har gjort i oppgaven, og denne er $1.21\cdot 10^{-13}$, som ville ført til at vi forkastet nullhypotesen og ville tro at flere enn 5 transporter var kjørt gjennom bykjernen. Men, dette var ikke med i oppgaven.

Oppgave 3 Trykkfasthet av murblokker

I denne oppgave er Y normalfordelt med $\mu = \mathrm{E}(Y)$, gitt i MPa (10^6 Pascal), og standardavvik $\sigma = \mathrm{SD}(Y) = 0.21$ MPa.

a) $Y \sim N(2.10, 0.21^2)$.

Hva er sannsynligheten for at en tilfeldig valgt murblokk har en trykkfasthet som er høyere enn 1.83 MPa, dvs. $P(Y>1.83)$?

$$\begin{aligned} P(Y>1.83) &= 1-P(Y\le 1.83) = 1-P(\frac{Y-2.10}{0.21} \le \frac{1.83-2.10}{0.21}) = 1-P(Z\le -1.29) \\ &= 1-\Phi(-1.29) = 1-0.0985 = \underline{\underline{0.9015}} \end{aligned}$$

Hva er sannsynligheten for at en tilfeldig valgt murblokk har en trykkfasthet som avviker mindre enn 0.3 MPa fra forventningsverdien $\mu = 2.10$ MPa?

$$\begin{aligned} P(|Y-\mu|<0.2) &= P(-0.3<Y-\mu\le 0.3) = P(Y-\mu\le 0.3) - P(Y-\mu\le -0.3) \\ &= P(\frac{Y-2.10}{0.21}\le\frac{0.3}{0.21}) - P(\frac{Y-2.10}{0.21}\le\frac{-0.3}{0.21}) \\ &= \Phi(1.43)-\Phi(-1.43) = 0.9236-0.0764 = \underline{\underline{0.8472}} \end{aligned}$$

Vi ser på måling av trykkfasthet for $n=24$ tilfeldig valgte murblokker fra produksjonen. Hva er sannsynligheten for at den minste målingen vil være lavere enn 1.83 MPa?

$$\begin{aligned} P(Y_{min}\le 1.83) &= 1-P(Y_{min}>1.83) = 1-P(Y_1>1.83\cap Y_2>1.83\cap\cdots\cap Y_{24}>1.83) \\ &= 1-[P(Y>1.83)]^{24} = 1-[1-P(Y\le 1.83)]^{24} = 1-(0.9015)^{24} = \underline{\underline{0.917}} \end{aligned}$$

Der overgangen fra første til andre linje er på grunn av uavhengighet i målt trykkfasthet mellom tilfeldig valgte murblokker.

b) Bedriften ønsker å undersøke om det er grunn til å tro at forventet trykkfasthet for den nye typen murblokker er lavere enn 2.40 MPa.

Null- og alternativ hypotese:

$$H_0: \mu = 2.40 \qquad H_1: \mu < 2.40$$

Her ser vi på parameteren μ som ukjent og parameteren $\sigma = 0.21$ er kjent.

Anta at vi har målt trykkfasthet til n tilfeldig valgte murblokker, og kall disse $Y_1, ..., Y_n$. Vi setter vi opp følgende estimator for μ.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$$

Vi vet at under H_0 så er

$$Z_0 = \frac{(\overline{Y}-2.40)}{\sigma\cdot\sqrt{\frac{1}{n}}} \quad \text{standard normalfordelt } N(0,1).$$

Vi vil forkaste H_0 når $Z_0 \le k$, der konstanten k finnes slik at Type-I feilen er kontrollert på nivå α.

$$\begin{aligned} P(Z_0\le k|H_0 \text{ sann}) &\le \alpha \\ k &\le -z_\alpha \end{aligned}$$

der z_α er α-kvantilen i en standard normalfordeling.

Forkastningsområde: $\underline{\underline{\text{Forkast } H_0 \text{ når } Z_0 \leq -z_\alpha}}$. Alternativt kan vi løse ut for $\overline{Y}$ og får heller regelen: Forkast H_0 når $\overline{Y} \leq 2.40 + z_\alpha \sigma \cdot \sqrt{\frac{1}{n}}$, innsatt $\sigma = 0.21$.

Når $\alpha = 0.05$ er $z_{0.05} = 1.645$, og i oppgaven er det oppgitt at $n = 24$ og $\overline{y} = 2.30$.

Dermed er $z_0 = \frac{\overline{y} - 2.40}{\sigma \cdot \sqrt{\frac{1}{n}}} = \frac{2.30 - 2.40}{0.21\sqrt{\frac{1}{24}}} = -2.33$.

Siden $z_0 = -2.33 < -z_{0.05} = -1.645$ så $\underline{\underline{\text{forkaster vi } H_0 \text{ på nivå } \alpha = 0.05}}$, og konkluderer med at trykkfastheten er mindre enn 2.40MPa.

Nå vil vi se hvor lett det er å forkaste H_0 med regelen vår hvis i virkeligheten $\mu = 2.30$MPa. 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). Dette betegnes *teststyrken*.

$$\begin{aligned}
P(\text{Forkaste } H_0 | H_0 \text{ gal og } \mu = 2.30) &= P(Z_0 < -z_\alpha | \mu = 2.30) \\
&= P(\overline{Y} \leq 2.40 + z_\alpha \sigma \cdot \sqrt{\frac{1}{n}} | \mu = 2.30) \\
&= P(\frac{\bar{Y} - 2.30}{\frac{\sigma}{\sqrt{n}}} < \frac{2.40 - 2.30}{\frac{\sigma}{\sqrt{n}}} - z_\alpha | \mu = 2.30) \\
&= \Phi(\frac{2.40 - 2.30}{\frac{\sigma}{\sqrt{n}}} - z_\alpha | \mu = 2.30) \\
&= \Phi(0.69) = \underline{\underline{0.7549}}
\end{aligned}$$

Oppgave 4 Hubble

a) Minste kvadraters metode minimerer $\text{SSE}(\beta) = \sum_{i=1}^{11} (y_i - \beta x_i)^2$.

$$\frac{d\text{SSE}}{d\beta} = 0$$

$$\sum_{i=1}^{11} y_i x_i - \beta \sum_{i=1}^{11} x_i^2 = 0$$

Dette tilsvarer: $\sum_{i=1}^{11} y_i x_i = \beta \sum_{i=1}^{11} x_i^2$ som gir svaret.
Innsetting gir $\hat{\beta} = 0.0567$.

Forventning og varians blir

$$E[\hat{\beta}] = \frac{\sum_{i=1}^{11} x_i E[Y_i]}{\sum_{i=1}^{11} x_i^2} = \frac{\sum_{i=1}^{11} x_i^2 \beta}{\sum_{i=1}^{11} x_i^2} = \beta$$

$$Var[\hat{\beta}] = \frac{\sum_{i=1}^{11} x_i^2 Var[Y_i]}{(\sum_{i=1}^{11} x_i^2)^2} = \frac{\sum_{i=1}^{11} x_i^2 \sigma^2}{(\sum_{i=1}^{11} x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{11} x_i^2}$$

b) Predikert verdi er $\hat{y}_0 = x_0 \hat{\beta} = 900 \cdot 0.0567 = 51.03$.

Vi har at $\hat{y}_0 - Y_0 = \hat{\beta} x_o - \beta x_0 - \epsilon_0 = x_0(\hat{\beta} - \beta) - \epsilon_0$, dvs $E[\hat{y}_0 - Y_0] = E[x_0(\hat{\beta} - \beta)] = 0$, og

$$Var[\hat{y}_0 - Y_0] = Var[x_0(\hat{\beta} - \beta) - \epsilon_0] = \frac{x_0^2 \sigma^2}{\sum_{i=1}^{11} x_i^2} + \sigma^2.$$

Et estimat for σ er $s = \sqrt{\frac{1}{10} 9.87} = 0.993$. Vi har at $T = \frac{\hat{y}_0 - Y_0}{s\sqrt{1 + \frac{900^2}{\sum_{i=1}^{11} x_i^2}}} \sim t_{10}$. Da blir et 95 prediksjonsintervall for observasjon Y_0 gitt ved

$$(\hat{y}_0 - t_{10,0.025} s \sqrt{1 + \frac{900^2}{\sum_{i=1}^{11} x_i^2}}, \hat{y}_0 + t_{10,0.025} s \sqrt{1 + \frac{900^2}{\sum_{i=1}^{11} x_i^2}}),$$

der $t_{0.025,10} = 2.23$.
Innsetting gir $(48.5, 53.5)$.