Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

LØSNINGSFORSLAG TIL EKSAMEN I FAG TMA4245 STATISTIKK

Lørdag 11.juni 2005

Oppgave 1 Meningsmålinger

a) Antagelser for at X er binomisk fordelt:

- Gjør n forsøk: Spør n personer.
- Registrerer suksess eller fiasko i hvert forsøk: Får svaret JA eller ikke JA (nei eller vet ikke) i hvert forsøk.
- P(suksess) lik i alle forsøk: Sannsynlighet for JA er p for alle som blir spurt.
- Forsøka er uavhengige: Rimelig å anta at de som blir spurt svarer uavhengig av hverandre.

$P(X \geq 18) = 1 - P(X < 18) = 1 - P(X \leq 17) \overset{\text{tabell}}{=} 1 - 0.965 = 0.035.$

$P(10 < X < 15) = P(X \leq 14) - P(X \leq 10) \overset{\text{tabell}}{=} 0.584 - 0.048 = 0.536$

b)
- $E(\hat{P}) = p$ og $\mathrm{Var}(\hat{P}) = \frac{1}{4}(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)$.
- $E(P^*) = p$ og $\mathrm{Var}(P^*) = \frac{1}{n_1+n_2}p(1-p)$.

Egenskaper for god estimator: forventningsrett og liten varians. Begge estimatorene er forventningsrette, men P^* har minst varians, vi velger derfor P^*.

La $\alpha = 0.05$. Siden $\frac{\hat{P}-p}{\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}}$ er tilnærmet standardnormalfordelt får vi:

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{P}-p}{\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}} < z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$P\left(\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})} < p < \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{2n}\hat{P}(1-\hat{P})}\right) \approx 1-\alpha$$

Et tilnærmet 95% konfidensintervall for p blir da:

$$\left[\hat{p} - z_{0.025}\sqrt{\frac{1}{2n}\hat{p}(1-\hat{p})}, \hat{p} + z_{0.025}\sqrt{\frac{1}{2n}\hat{p}(1-\hat{p})}\right].$$

c) Vi har at

$$Y = X_3 - n\hat{P} = X_3 - n\frac{X_1 + X_2}{2n} = X_3 - \frac{1}{2}X_1 - \frac{1}{2}X_2.$$

Siden n er stor og p ikke nær 0 og 1, vil vi ha at $np > 5$ og $n(1-p) > 5$, slik at vi kan bruke normaltilnærming til binomisk fordeling. Vi kan dermed anta at X_1, X_2 og X_3 alle er tilnærmet normalfordelt, de er uavhengige, og lineærkombinasjonen Y er dermed også tilnærmet normalfordelt.

$\text{Var}(Y) = \text{Var}(X_3 - n\hat{P}) \overset{\text{uavh.}}{=} \text{Var}(X_3) + n^2\text{Var}(\hat{P}) \overset{b)}{=} np(1-p) + n^2\frac{1}{2n}p(1-p) = \frac{3}{2}np(1-p).$

Har da at

- $X_3 - n\hat{P}$ er tilnærmet normalfordelt
- $\text{Var}(X_3 - n\hat{P}) = \frac{3}{2}np(1-p)$
- $\text{E}(X_3 - n\hat{P}) = E(X_3) - nE(\hat{P}) = np - np = 0$

Vi får da et prediksjonsintervall ved:

$$P\left(-z_{\frac{\alpha}{2}} < \frac{X_3 - n\hat{P}}{\sqrt{\frac{3}{2}np(1-p)}} < z_{\frac{\alpha}{2}}\right) \approx 1-\alpha$$

$$P\left(n\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{3}{2}np(1-p)} < X_3 < n\hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{3}{2}np(1-p)}\right) \approx 1-\alpha$$

Siden n er stor, vil variansen til $\hat{P}$ være liten, og $\hat{P}$ være en god estimator for p. Vi kan derfor erstatte p med estimatet $\hat{p}$ i uttrykket for intervallgrensene.

Intervallet blir: $[n\hat{p} - z_{0.025}\sqrt{\frac{3}{2}n\hat{p}(1-\hat{p})}, n\hat{p} + z_{0.025}\sqrt{\frac{3}{2}n\hat{p}(1-\hat{p})}]$

Innsatt verdier blir intervallet $[633, 704]$.

Oppgave 2 Veiprosjektet

a) Vi jobber med X som er normalfordelt med forventning $\mu = 10000$ kr/meter og standardavvik $\sigma = 2500$ kr/meter.

$$\begin{aligned} P(X > 13000) &= 1 - P(X \leq 13000) = 1 - P(\frac{X - 13000}{2500} \leq \frac{13000 - 10000}{2500}) = 1 - P(Z \leq 1.2) \\ &= 1 - \Phi(1.2) = 1 - 0.8848 = \underline{\underline{0.1152}} \end{aligned}$$

Finn et tall, k, slik at sannsynligheten er 0.05 for at konstnaden pr. meter for veien vil bli mindre enn k.

$$\begin{aligned} P(X < k) &= 0.05 \\ P(\frac{X - 10000}{2500} < \frac{k - 10000}{2500}) &= 0.05 \\ \frac{k - 10000}{2500} &= -1.645 \\ k &= -1.645 \cdot 2500 + 10000 = \underline{\underline{5887.5}} \end{aligned}$$

Gitt at vi vet at kostnaden pr. meter blir minst 10000 kr/meter, hva er da sannsynligheten for at kostnaden pr. meter blir høyere enn 13000 kr/meter?

$$\begin{aligned} P(X > 13000|X > 10000) &= \frac{P(X > 13000 \cap X > 10000)}{P(X > 10000)} \\ &= \frac{P(X > 13000)}{P(X > 10000)} = 2 \cdot 0.1152 = 0.23 \end{aligned}$$

b) Null- og alternativ hypotese:

$H_0 : \mu = 10000 \qquad H_1 : \mu > 10000$

De ukjente parameterene er μ og σ^2, og vi setter opp følgende estimatorer:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X} \\ S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \end{aligned}$$

Vi vet at under H_0 så er

$$T_0 = \frac{(\overline{X} - 10000)}{S \cdot \sqrt{\frac{1}{n}}} \quad t\text{-fordelt med } (n-1) \text{ frihetsgrader.}$$

Vi vil forkaste H_0 når $T_0 \geq k$, der konstanten k finnes slik at Type-I feilen er kontrollert på nivå α.

$$\begin{aligned} P(T_0 \geq k | H_0 \text{ sann}) &\leq \alpha \\ k &\leq t_{\alpha,(n-1)} \end{aligned}$$

der $t_{\alpha,(n-1)}$ er α-kvantilen i en t-fordeling med $n-1$ frihetsgrader.

Forkastningsmråde: $\underline{\underline{\text{Forkast } H_0 \text{ når } T_0 \geq t_{\alpha,(n-1)}}}$.

Når $\alpha = 0.01$ og $n = 9$ er $t_{0.01,8} = 2.896$. Innsatt data fra tabell 1 i oppgaveteksten har vi:

$$\begin{aligned} \overline{x} &= \frac{106480}{9} = 11831.11 \\ s^2 &= \frac{1}{8}\sum_{i=1}^{9}(x_i - \overline{x})^2 = \frac{49295335}{8} = 6161917 \\ s &= \sqrt{6161917} = 2482.3 \\ t_0 &= \frac{11831.11 - 10000}{\frac{2482.3}{\sqrt{9}}} = 2.21 \end{aligned}$$

Siden $t_0 = 2.21 < t_{0.01,8} = 2.896$ så $\underline{\underline{\text{forkaster vi } \textit{ikke } H_0 \text{ på nivå } \alpha = 0.01}}$, og konkluderer med at vi har ikke tilstrekkelig bevis til å anta at kostnadene blir større enn 10000 kr pr. meter.

Når vi ikke forkastet H_0 på nivå 0.01 så betyr det at p-verdien må være større enn 0.01. P-verdien er gitt som

$$P(T_0 > t_0 | H_0 \text{ sann}) = P(T_0 > 2.21 | \mu = 10000) = 1 - P(T_0 \leq 2.21 | \mu = 10000)$$

der T_0 under H_0 er t-fordelt med $n-1$ frihetsgrader. Fra tabell 2 i oppgaven så slår vi opp på $P(T \leq t)$ med $t = 2.2$ og $\nu = 8$, og finner 0.971, som gir p-verdi $1 - 0.071 = \underline{\underline{0.029}}$.

c) La $g(x)$ være sannsynlighetstettheten til X, og $h(y)$ være sannsynlighetstettheten til Y.

Siden X og Y er uavhengige, er simultan sannsynlighetstetthet $f(x,y) = g(x) \cdot h(y)$.

Forventingen til $W = X \cdot Y$:

$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(X \cdot Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot y \cdot g(x) \cdot h(y) dx dy \\ &= \int_{-\infty}^{\infty} x \cdot g(x) dx \cdot \int_{-\infty}^{\infty} y \cdot h(y) dy = \mathrm{E}(X) \cdot \mathrm{E}(Y) = \mu \cdot \eta \end{aligned}$$

Alternativt: $\mathrm{Cov}(X,Y) = \mathrm{E}(X \cdot Y) - \mathrm{E}(X) \cdot \mathrm{E}(Y) = 0$ når X og Y er uavhengige, slik at $\mathrm{E}(W) = \mathrm{E}(X \cdot Y) = \mathrm{E}(X) \cdot \mathrm{E}(Y) = \mu \cdot \eta$.

Før vi begynner på variansen, trenger vi følgende sammenhenger:

$$\begin{aligned}
\mathrm{E}(W^2) &= \mathrm{E}(X^2 \cdot Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 \cdot y^2 \cdot g(x) \cdot h(y) dx dy \\
&= \int_{-\infty}^{\infty} x^2 \cdot g(x) dx \cdot \int_{-\infty}^{\infty} y^2 \cdot h(y) dy = \mathrm{E}(X^2) \cdot \mathrm{E}(Y^2)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}(X^2) - \mu^2 \\
\mathrm{E}(X^2) &= \mathrm{Var}(X) + \mu^2 = \sigma^2 + \mu^2 \\
\mathrm{E}(Y^2) &= \mathrm{Var}(Y) + \eta^2 = \tau^2 + \eta^2
\end{aligned}$$

Variansen til $W = X \cdot Y$:

$$\begin{aligned}
\mathrm{Var}(W) &= \mathrm{E}(W^2) - [\mathrm{E}(W)]^2 = \mathrm{E}(X^2) \cdot \mathrm{E}(Y^2) - [\mathrm{E}(X) \cdot \mathrm{E}(Y)]^2 \\
&= (\sigma^2 + \mu^2) \cdot (\tau^2 \cdot \eta^2) - [\mu \cdot \eta]^2 \\
&= \sigma^2 \cdot \tau^2 + \sigma^2 \cdot \eta^2 + \tau^2 \cdot \mu^2 + \mu^2 \cdot \eta^2 - \mu^2 \cdot \eta^2 \\
&= \sigma^2 \cdot \tau^2 + \sigma^2 \cdot \eta^2 + \tau^2 \cdot \mu^2
\end{aligned}$$

Oppgave 3 Kalibrering ved regresjon

a) Y_i, $i = 1, \ldots, m$ er uavhengige normalfordelte variabler, dermed er den lineære kombinasjonen $\overline{Y}$ også normalfordelt. Siden $\mathrm{E}[\overline{Y}] = E[Y] = \alpha$ $(x = 0)$ og $\mathrm{Var}(\overline{Y}) = \mathrm{Var}(Y)/m = \sigma^2/m$, blir fordelingen til $\overline{Y}$ lik $n(\cdot; \alpha, \sigma/\sqrt{m})$.

Har sett at $\mathrm{E}[\overline{Y}] = \alpha$, så $\overline{Y}$ er en forventningsrett estimator for α. Siden $\mathrm{Var}(\overline{Y}) = \sigma^2/m \to 0$ når $m \to \infty$, følger det at mer data gir skarpere estimat, som er et rimelig krav til en estimator.

b) Ut fra betraktningen over, får vi følgende rimelighetsfunksjon (likelihood-funksjon)

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\}$$

Bestemmer estimatoren ut fra ligningen

$$\frac{\partial}{\partial \beta} \ln L(\beta) = \frac{\partial}{\partial \beta}\big(- n \ln(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\big) = 0$$

Det gir ligningen

$$\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0$$

(og siden $\partial^2 \ln L(\beta)/\partial\beta^2 = -\sum_{i=1}^{n} x_i^2 < 0$ svarer løsningen til et maks.punkt).
Dermed blir punktestimatet for SME gitt ved

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - \alpha)}{\sum_{i=1}^{n} x_i^2},$$

og SME blir

$$B = \frac{\sum_{i=1}^{n} x_i(Y_i - \alpha)}{\sum_{i=1}^{n} x_i^2}$$

Minste kvadraters estimator fås ved å bestemme den verdien for β som minimerer

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

som igjen gir ligningen

$$\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0$$

dvs. den samme som ovenfor, og vi får samme esitmator som SME.