TMA4245 Statistikk
Vår 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 11, blokk II
Løsningsskisse

Oppgave 1

a) Merk fra Venn diagram at I ikke overlapper F eller R.

$$P(R|F) = \frac{P(R \cap F)}{P(F)} = \frac{0.3}{0.5} = 0.6$$

$$P(R|I') = \frac{P(R \cap I')}{P(I')} = \frac{P(R)}{1 - P(I)} = \frac{0.4}{1 - 0.05} = 0.421$$

b) Generelle forutsetninger for binomisk fordeling
i) Forsøksrekken består av n enkeltforsøk.
ii) Det registreres kun suksess eller ikke suksess.
iii) Sannsynligheten for suksess er lik i alle forsøk.
iv) Enkeltforsøkene er uavhengige.

For X har vi
i) Det er valgt ut n kamper.
ii) Vi registrerer kun om den som får første målet vinner(suksess) eller ikke.
iii) Sannsynligheten for suksess er p og er antatt å være konstant.
iv) Vi antar at kampene er uavhengige.
Dette er rimelige antakelser.

Sentralgrenseteoremet sier:
Dersom $Z_1, Z_2, \ldots, Z_n$ er uavhengig identisk fordelte fra sannsynlighetsfordelingen $f_Z(z)$, hvor $E(Z) = \mu$ og $Var(Z) = \sigma^2$, så vil $\sqrt{n}\frac{\overline{Z}-\mu}{\sigma}$ konvergere mot en normalfordeling med forventning 0 og varians 1. Der $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

For en binomisk forsøksrekke, definer Z_i slik at: $Z_i = 1$ hvis suksess, og $Z_i = 0$ ellers. Med andre ord:

$$P(Z_i = z) = \begin{cases} p & \text{hvis } z = 1 \\ 1 - p & \text{hvis } z = 0 \end{cases}$$

Slik at $E(Z_i) = p$ og $Var(Z_i) = p(1 - p)$.

Siden enkeltforsøkene er uavhengige så er Z_i ene også uavhengige. Av sentralgrenseteoremet følger at $\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}}$ konvergerer mot en normalfordeling med forventning 0 og varians 1. Der $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

c) H_0: $p \geq 0.8$ mot H_1: $p < 0.8$
Eventuelt: H_0: $p = 0.8$ mot H_1: $p < 0.8$

Vi ønsker å forkaste dersom $\hat{p} < k$, hvor k bestemmes slik at

$$P(\hat{p} < k) = \alpha = 0.05$$

Vi benytter at $Z = \sqrt{n}\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)}}$ er tilnærmet normalfordelt med forventing 0 og varians 1 under H_0. Da har vi fra ligningen over:

$$P(Z < \frac{\sqrt{n}(k-p_0)}{\sqrt{p_0(1-p_0)}}) = 0.05$$

$$\frac{\sqrt{n}(k-p_0)}{\sqrt{p_0(1-p_0)}} = -Z_{0.05}$$

Dette gir $k = p_0 - Z_{0.05}\sqrt{\frac{p_0(1-p_0)}{n}}$. Vi forkaster H_0 dersom:

$$\hat{p} < p_0 - Z_{0.05}\sqrt{\frac{p_0(1-p_0)}{n}} = 0.8 - 0.658\frac{1}{\sqrt{n}}$$

For $n = 24$ og $X = \sum_{i=1}^{n} Z_i = 17$ får vi $\hat{p} = 0.71$, $k = 0.67$. Vi forkaster ikke H_0.

Vi kan ikke pastå at ekspertkommentatoren tar feil på 5 prosent nivå.

d) Vi ønsker at styrken på testen i alternativet $p = 0.7$ skal være minst 0.9. Dvs

$$P(\hat{p} < 0.8 - 0.658\frac{1}{\sqrt{n}}|p = 0.7) = 0.9$$

Vi benytter at $Z = \sqrt{n}\frac{\hat{p}-0.7}{\sqrt{0.7\cdot 0.3}}$ er tilnærmet normalfordelt med forventing 0 og varians 1 under alternativet med $p = 0.7$. Innsatt i kravet fra ligningen over gir dette:

$$P(Z < \sqrt{n}\frac{0.8-0.7}{\sqrt{0.7\cdot 0.3}} - \frac{0.658}{\sqrt{0.7\cdot 0.3}}) = 0.9$$

0.1 percentilen i normalfordelingen er lik $Z_{0.1} = 1.28$. Kravet som n må oppfylle blir dermed:

$$\sqrt{n}\frac{0.1}{\sqrt{0.7\cdot 0.3}} - \frac{0.658}{\sqrt{0.7\cdot 0.3}} = 1.28$$

Løsningen blir $n = 155.1$ kamper. Dvs at vi må se minst 156 kamper for å oppnå den ønskede styrken på testen.

Oppgave 2

a) For at X skal være tilnærmet binomisk fordelt må en ha at $N >> n$ slik at det ikke betyr noe at samme person ikke vil bli spurt mer enn en gang.

$$P(X=9) = \binom{n}{9}\theta^9(1-\theta)^{n-9} = \binom{20}{9}0.5^9(1-0.5)^{20-9} = \underline{\underline{0.1602}}$$

$$P(X>9) = 1 - P(X \leq 9) \overset{\text{tabell}}{=} 1 - 0.412 = \underline{\underline{0.588}}$$

$$\begin{aligned} P(X>9|X\leq 12) &= \frac{P(X>9 \cap X \leq 12)}{P(X\leq 12)} = \frac{P(10 \leq X \leq 12)}{P(X \leq 12)} \\ &= \frac{P(X\leq 12) - P(X \leq 9)}{P(X\leq 12)} \overset{\text{tabell}}{=} \frac{0.868-0.412}{0.868} = \underline{\underline{0.525}} \end{aligned}$$

b) $X \sim b(x;n,\theta)$

$$L(\theta) = P(X=x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$
$$l(\theta) = ln\binom{n}{x} + xln\theta + (n-x)ln(1-\theta)$$
$$l'(\theta) = x\frac{1}{\theta} + (n-x)\frac{1}{1-\theta}(-1) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

$$\begin{gathered} l'(\theta) = 0 \\ \Downarrow \\ \frac{x}{\theta} = \frac{n-x}{1-\theta} \\ x(1-\theta) = \theta(n-x) \\ x - x\theta = \theta n - \theta x \\ \theta = \frac{x}{n} \end{gathered}$$

dvs. SME er $\underline{\underline{\hat{\theta} = \frac{X}{n}}}$

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathrm{E}(X) = \frac{1}{n}n\theta = \underline{\underline{\theta}}$$

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{1}{n^2}n\theta(1-\theta) = \underline{\underline{\frac{\theta(1-\theta)}{n}}}$$

c) $H_0 : \theta = \frac{1}{2}$ mot $H_1 : \theta \neq \frac{1}{2}$

Benytter testobservator

$$Z = \frac{X - n\frac{1}{2}}{\sqrt{n\frac{1}{2}(1-\frac{1}{2})}} \approx N(z;0,1)$$

når H_0 er riktig.

Dvs. forkaster H_0 dersom

$$Z < -z_{\frac{\alpha}{2}} \quad \text{eller} \quad Z > z_{\frac{\alpha}{2}}$$
$$\Updownarrow$$
$$\frac{X - \frac{n}{2}}{\frac{\sqrt{n}}{2}} < -z_{\frac{\alpha}{2}} \quad \text{eller} \quad \frac{X - \frac{n}{2}}{\frac{\sqrt{n}}{2}} > z_{\frac{\alpha}{2}}$$

Innsatt tall:

Observert verdi av testobservator:

$$Z = \frac{2562 - \frac{5000}{2}}{\frac{\sqrt{5000}}{2}} = 1.75$$

$\alpha = 0.10$ gir $z_{\frac{\alpha}{2}} = 1.645$

Dvs. forkaster H_0 og erklærer G som vinner av valget.

Oppgave 3

a) La $u_i = x_i(1 - x_i)$, $i = 1, \ldots, n$. Modellen for avviket Y_i blir da

$$Y_i = au_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Likelihood-funksjonen er sannsynlighetstettheten til $y_1, \ldots, y_n$ gitt $x_1, \ldots, x_n$ med parameter a, som blir

$$\begin{aligned} L(a) &= f(y_1, \ldots, x_n | x_1, \ldots, x_n, a) = f(y_1|x_1, a) \cdots f(y_n|x_n, a) \\ &= \prod_{i=1}^{n} f(y_i|x_i, a) = \prod_{i=1}^{n} n(y_i; au_i, \sigma_0^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(y_i - au_i)^2}. \end{aligned}$$

Videre blir log-likelihoodet

$$\begin{aligned} \ell(a) = \log L(a) &= \sum_{i=1}^{n} \left\{ -\frac{1}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}(y_i - au_i)^2 \right\} \\ &= -\frac{n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - au_i)^2. \end{aligned}$$

Derivasjon med hensyn på a gir

$$\begin{aligned} \frac{\partial \ell(a)}{\partial a} &= -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} 2(y_i - au_i)(-u_i) \\ &= \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (y_i u_i - au_i^2) \\ &= \frac{1}{\sigma_0^2} \sum_{i=1}^{n} y_i u_i - \frac{a}{\sigma_0^2} \sum_{i=1}^{n} u_i^2. \end{aligned}$$

Verdien av a som maksimerer $\ell(a)$, og dermed også $L(a)$, finnes ved å sette den deriverte lik null, og løse for a,

$$\begin{aligned}
\frac{\partial \ell(a)}{\partial a} &= 0 \\
\frac{a}{\sigma_0^2}\sum_{i=1}^{n} u_i^2 &= \frac{1}{\sigma_0^2}\sum_{i=1}^{n} y_i u_i \\
a &= \frac{\sum_{i=1}^n y_i u_i}{\sum_{i=1}^n u_i^2} = \frac{\sum_{i=1}^n x_i(1-x_i)y_i}{\sum_{i=1}^n (x_i(1-x_i))^2}.
\end{aligned}$$

Det følger at sannsynlighetsmaksimeringsestimatoren for a er

$$\hat{a} = \frac{\sum_{i=1}^n x_i(1-x_i)Y_i}{\sum_{i=1}^n (x_i(1-x_i))^2}.$$

b) Foventningsverdien til estimatoren er

$$\begin{aligned}
\mathrm{E}(\hat{a}) &= \mathrm{E}\left(\frac{\sum_{i=1}^n u_i Y_i}{\sum_{i=1}^n u_i^2}\right) \\
&= \frac{\sum_{i=1}^n u_i \mathrm{E}(Y_i)}{\sum_{i=1}^n u_i^2} \\
&= \frac{\sum_{i=1}^n u_i \cdot a u_i}{\sum_{i=1}^n u_i^2} \\
&= a\frac{\sum_{i=1}^n u_i^2}{\sum_{i=1}^n u_i^2} = a,
\end{aligned}$$

og variansen er

$$\begin{aligned}
\mathrm{Var}(\hat{a}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n y_i u_i}{\sum_{i=1}^n u_i^2}\right) \\
&= \frac{1}{\left(\sum_{i=1}^n u_i^2\right)^2}\sum_{i=1}^{n} u^2 \mathrm{Var}(Y_i) \\
&= \frac{\sum_{i=1}^n u^2\sigma_0^2}{\left(\sum_{i=1}^n u_i^2\right)^2} \\
&= \frac{\sigma_0^2}{\sum_{i=1}^n u_i^2}
\end{aligned}$$

Siden observasjonene Y_i, $i = 1, \ldots, n$ er normalfordelte, og estimatoren $\hat{a}$ er en lineærkombinasjon av $Y_1, \ldots, Y_n$, vet vi at $\hat{a}$ vil være normalfordelt med forventningsverdi $\mathrm{E}(\hat{a})$ og varians $\mathrm{Var}(\hat{a})$ som over,

$$\hat{a} \sim N\left(a, \sigma_0^2 / \sum_{i=1}^{n} u_i^2\right).$$

c) Siden a vil være lik null for en ideell blanding, og avvike fra null for en ikke-ideell blanding, kan følgende hypoteser brukes:

$$H_0: \; a = 0, \qquad H_1: \; a \neq 0.$$

Under nullhypotesen har vi

$$\hat{a} \sim N\left(0, \mathrm{Var}(\hat{a})\right),$$

slik at observatoren

$$Z = \frac{\hat{a}}{\sqrt{\mathrm{Var}(\hat{a})}}$$

er standard normalfordelt. Vi forkaster H_0 på signifikansnivå α dersom $|Z| > z_{\alpha/2}$, som er ekvivalent med at $|\hat{a}| > z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{a})}$. Med $\alpha = 5\%$ får vi $z_{\alpha/2} = 1.96$. Variansen $\sigma_0^2 = 0.025^2$ og kvadratsummene $\sum_{i=1}^n u_i y_i = -0.0945$ og $\sum_{i=1}^n u_i^2 = 0.3333$ er gitt i oppgaveteksten. Ved å sette inn disse tallverdiene finner vi estimatet

$$\hat{a} = \frac{\sum_{i=1}^n u_i y_i}{\sum_{i=1}^n u_i^2} = \frac{-0.0945}{0.3333} = -0.2835,$$

og den kritiske verdien

$$z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{a})} = z_{0.025}\sqrt{\frac{\sigma_0^2}{\sum_{i=1}^n u_i^2}} = 1.96\frac{0.025}{\sqrt{0.3333}} = 0.0849.$$

Det at $\hat{a} < -z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{a})}$ betyr at $|Z| > z_{\alpha/2}$, så vi forkaster H_0 på signifikansnivå $\alpha = 5\%$. Det er med andre ord grunnlag for å hevde at blandingen ikke er ideell.

d) Stigningstallet til tangentlinja er

$$\frac{\mathrm{E}(Y|x) - z(x)}{x - 0} = \frac{ax(1-x) - z(x)}{x}.$$

Dette skal være lik den deriverte til $\mathrm{E}(Y|x) = ax(1-x)$, som er

$$\frac{\partial}{\partial x} ax(1-x) = a(1-2x).$$

Ved å sette uttrykkene lik hverandre, får vi

$$\begin{aligned}
\frac{ax(1-x) - z(x)}{x} &= a(1-2x) \\
ax(1-x) - z(x) &= ax(1-2x) \\
z(x) &= ax(1-x) - ax(1-2x) = ax^2.
\end{aligned}$$

En rimelig estimator for $z(x)$ er dermed $\hat{z}(x) = \hat{a}x^2$.

Et 95% konfidensintervall for a er

$$\hat{a} \pm z_{0.025}\sqrt{\mathrm{Var}(\hat{a})} = \hat{a} \pm 1.96\frac{\sigma_0}{\sqrt{\sum_{i=1}^n u_i^2}}.$$

Et tilsvarende 95% konfidensintervall for $z(x_0) = ax_0^2$ er

$$\hat{a}x_0^2 \pm z_{0.025}\sqrt{\mathrm{Var}(\hat{a}x_0^2)} = \hat{a}x_0^2 \pm 1.96\sqrt{\frac{\sigma_0^2 \cdot (x_0^2)^2}{\sum_{i=1}^n u_i^2}} = \hat{a}x_0^2 \pm 1.96\frac{\sigma_0 \cdot x_0^2}{\sqrt{\sum_{i=1}^n u_i^2}}.$$

Grensene til konfidensintervallet for $z(x_0)$ er altså lik grensene for intervallet for a, multiplisert med x_0^2.

Med $x_0 = 2/3$ blir konfidensintervallet for a

$$(-0.3684, -0.1987),$$

mens konfidensintervallet for $z(x_0)$ blir

$$(-0.3684 \cdot (2/3)^2, -0.1987 \cdot (2/3)^2) \quad \text{eller} \quad (-0.1637, -0.0883).$$