Løsningsforslag til eksamen i TMA4245 Statistikk 7. juni 2007

Oppgave 1: Pengespill

a)

For hver deltaker har vi følgende situasjon:

- Deltakeren får en serie oppgaver.
- Hver runde har to mulige utfall: Deltakeren klarer ikke oppgaven og går ut av konkurransen (hendelse A), eller han/hun klarer oppgaven og går videre til neste runde (hendelse A').
- Sannsynligheten for ikke å klare oppgaven, $p = P(A)$, er lik i hver runde.
- Resultatene fra hver runde er uavhengige.

Denne situasjonen svarer til en Bernoulli-forsøksrekke, der vi ikke bestemmer antall forsøk på forhånd, men repeterer forsøket (gir nye oppgaver) inntil første gang hendelsen A (klarer ikke oppgaven) inntreffer. Siden X er antall forsøk inntil A inntreffer *første* gang (deltakeren første gang ikke klarer oppgaven), er det rimelig å anta at X er geometrisk fordelt.

Sannsynligheten for at deltakeren går ut i første runde:

$$P(X=1) = f(1) = p(1-p)^{1-1} = p = \underline{\underline{0{,}10}}$$

Sannsynligheten for at deltakeren fortsatt er med etter fem runder:

$$P(X>5) = 1 - P(X \le 5) = 1 - F(5) = 1 - (1-(1-p)^5) = (1-p)^5 = 0{,}90^5 = \underline{\underline{0{,}59}}.$$

Sannsynligheten for at deltakeren ikke klarer oppgaven i niende runde ($X = 9$), dersom deltakeren klarer oppgavene til og med femte runde ($X > 5$): Her bruker vi betinget sannsynlighet, og resultatet fra forrige spørsmål.

$$\begin{aligned} P(X=9 \mid X>5) &= \frac{P(X=9 \cap X>5)}{P(X>5)} = \frac{P(X=9)}{P(X>5)} = \frac{f(9)}{1-F(5)} \\ &= \frac{p(1-p)^{9-1}}{(1-p)^5} = p(1-p)^3 = 0{,}10 \cdot 0{,}90^3 = \underline{\underline{0{,}073}} \end{aligned}$$

b)

Rimelighetsfunksjonen er

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

Tar logaritmen, deriverer og setter lik null:

$$\begin{aligned}
l(x_1,\ldots,x_n;p) &= n\ln(p) + (\sum_{i=1}^{n}(x_i-1))\ln(1-p) \\
\frac{d}{dp}l(x_1,\ldots,x_n;p) &= \frac{n}{p} + \frac{\sum_{i=1}^{n}x_i - n}{1-p}\cdot(-1) \\
&= \frac{n}{p} - \frac{\sum_{i=1}^{n}x_i - n}{1-p} \quad = \quad 0
\end{aligned}$$

Ved å multiplisere med $p(1-p)$ på begge sider får vi

$$\begin{aligned}
n(1-p) - (\sum_{i=1}^{n}x_i - n)p &= n - np - (\sum_{i=1}^{n}x_i)\,p + np \quad = \quad n - (\sum_{i=1}^{n}x_i)\,p \quad = \quad 0 \\
p &= \frac{n}{\sum_{i=1}^{n}x_i}
\end{aligned}$$

Sannsynlighetsmaksimeringsestimatoren for p blir

$$\underline{\underline{\hat{p} = \frac{n}{\sum_{i=1}^{n}X_i}}}$$

Med $n = 8$ og observerte antall runder som gitt i oppgaven, blir estimatet

$$\hat{p} = \frac{8}{\sum_{i=1}^{8}x_i} = \frac{8}{109} = \underline{\underline{0{,}073}}.$$

c)

Vi har følgende situasjon for hver oppgavelager:

- Resultater for et visst antall (n_1 eller n_2) deltakere blir registrert
- To mulig utfall: Deltakeren klarer færre enn fem oppgaver (hendelse C), eller ikke (dvs. klarer fem eller flere, hendelse C').
- Sannsynligheten for C er lik i for hver deltaker.

- Resultatene for hver deltaker er uavhengige.

Dette svarer til et binomisk forsøk, og Z_1 og Z_2 er dermed binomisk fordelte, med parametre som gitt i oppgaven.

Konfidensintervall for $q_1 - q_2$:

En rimelig estimator for $q_1 - q_2$ er $\hat{q}_1 - \hat{q}_2$. Vi finner først fordelingen til denne.

Siden vi kan anta at Z_1 og Z_2 er tilnærmet normalfordelte, er også $\hat{q}_1$ og $\hat{q}_2$ og dermed også $\hat{q}_1 - \hat{q}_2$ tilnærmet normalfordelte (alle disse tre estimatorene er lineærkombinasjoner av tilnærmet normalfordelte variabler).

Forventningsverdien til $\hat{q}_1 - \hat{q}_2$ er

$$\mathrm{E}(\hat{q}_1 - \hat{q}_2) = \mathrm{E}(\frac{Z_1}{n_1}) - \mathrm{E}(\frac{Z_1}{n_1}) = \frac{n_1 q_1}{n_1} - \frac{n_2 q_2}{n_2} = q_1 - q_2.$$

Variansen til $\hat{q}_1 - \hat{q}_2$ er

$$\begin{aligned}
\mathrm{Var}(\hat{q}_1 - \hat{q}_2) &\overset{uavh}{=} \mathrm{Var}(\frac{Z_1}{n_1}) + \mathrm{Var}(\frac{Z_1}{n_1}) = \frac{1}{n_1^2}\mathrm{Var}(Z_1) + \frac{1}{n_2^2}\mathrm{Var}(Z_2) \\
&= \frac{1}{n_1^2} n_1 q_1(1-q_1) + \frac{1}{n_2^2} n_2 q_2 (1-q_2) \\
&= \frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}.
\end{aligned}$$

Dermed er

$$Z = \frac{\hat{q}_1 - \hat{q}_2 - (q_1 - q_2)}{\sqrt{\frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}}}$$

tilnærmet standard normalfordelt.

For å lage konfidensintervall, bruker vi at:

$$P(-z_{0,05/2} < \frac{\hat{q}_1 - \hat{q}_2 - (q_1 - q_2)}{\sqrt{\frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}}} < z_{0,05/2}) \approx 0{,}95$$

Vi tilnærmer q_1 og q_2 i nevneren med $\hat{q}_1$ og $\hat{q}_2$ slik at

$$P(-z_{0,05/2} < \frac{\hat{q}_1 - \hat{q}_2 - (q_1 - q_2)}{\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}} < z_{0,05/2}) \approx 0{,}95$$

Vi løser ulikhetene slik at vi får $q_1 - q_2$ i midten, som gir

$$P\Big(\hat{q}_1 - \hat{q}_2 - z_{0,05/2}\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}} < q_1 - q_2 < \hat{q}_1 - \hat{q}_2 + z_{0,05/2}\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}\Big) \approx 0{,}95$$

Et tilnærmet 95% konfidensintervall for $q_1 - q_2$ blir

$$\underline{\underline{\left[\hat{q}_1 - \hat{q}_2 - z_{0,05/2}\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}, \hat{q}_1 - \hat{q}_2 + z_{0,05/2}\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}\right]}}$$

Innsatt verdier får vi intervallet $\underline{\underline{[0{,}08, 0{,}41]}}$.

Siden intervallet ikke inneholder 0, så gir det TV-selskapet grunn til å hevde at oppgavene har ulik vanskelighetsgrad.

Oppgave 2: Radar

a)

Vi benytter den kumulative fordelingsfunksjonen i oppgaveteksten. Regner først ut sannsynligheten for generell verdi av β, for så å regne ut for $\beta = \pi/8$. Dette gir

$$\mathrm{P}(Y > \pi/4) = 1 - \mathrm{P}(Y \leq \pi/4) = 1 - \frac{1 - \exp\left\{-\frac{\pi}{4\beta}\right\}}{1 - \exp\left\{-\frac{\pi}{2\beta}\right\}} = \frac{\exp\left\{-\frac{\pi}{4\beta}\right\} - \exp\left\{-\frac{\pi}{2\beta}\right\}}{1 - \exp\left\{-\frac{\pi}{2\beta}\right\}} = \underline{\underline{0{,}1192}}$$

$$\begin{aligned}\mathrm{P}(\pi/4 < Y < \pi/3) = \mathrm{P}(Y < \pi/3) - \mathrm{P}(Y < \pi/4) &= \frac{1 - \exp\left\{-\frac{\pi}{4\beta}\right\}}{1 - \exp\left\{-\frac{\pi}{2\beta}\right\}} - \frac{1 - \exp\left\{-\frac{\pi}{3\beta}\right\}}{1 - \exp\left\{-\frac{\pi}{2\beta}\right\}} \\ &= \frac{\exp\left\{-\frac{\pi}{4\beta}\right\} - \exp\left\{-\frac{\pi}{3\beta}\right\}}{1 - \exp\left\{-\frac{\pi}{2\beta}\right\}} = \underline{\underline{0{,}0671}}\end{aligned}$$

$$\begin{aligned}\mathrm{P}(Y > \pi/4 | Y < \pi/3) = \frac{\mathrm{P}(Y > \pi/4 \cap Y < \pi/3)}{\mathrm{P}(Y < \pi/3)} &= \frac{0{,}0671}{\left(1 - \exp\left\{-\frac{\pi}{3\beta}\right\}\right) / \left(1 - \exp\left\{-\frac{\pi}{2\beta}\right\}\right)} \\ &= \underline{\underline{0{,}0708}}.\end{aligned}$$

b)

Siden Y er en kontinuerlig, kan vi finne sannsynlighetstettheten ved å derivere den kumulative fordelingsfunksjonen i oppgaveteksten

$$f(y;\beta) = \frac{\mathrm{d}}{\mathrm{d}y}F(y;\beta) = \frac{1}{1-\exp\left\{-\frac{\pi}{2\beta}\right\}}\left(0-\left(-\frac{1}{\beta}\right)\exp\left\{-\frac{y}{\beta}\right\}\right) =$$

$$= \underline{\underline{\frac{1}{\beta-\beta\exp\left\{-\frac{\pi}{2\beta}\right\}}\exp\left\{-y/\beta\right\}}}$$

Fra figuren i oppgaveteksten har vi at $\tan(Y) = X$, altså har vi en-til-en relasjon mellom vinkelen Y og avstanden X. Det betyr at vi kan benytte transformasjon av variable (kap 7.2 i læreboka) til å finne fordelingen til X. La $y = \arctan(x) = w(x)$, altså den omvendte funksjonen av funksjonen over. Vi har da at sannsynlighetsfordelingen til X, $g(x;\beta)$, er gitt ved

$$g(x;\beta) = f(w(x);\beta)\cdot|w'(x)|,$$

Opplysningen i oppgaven eller oppslag i Rottmann gir at $w'(x) = 1/(1+x^2)$ som gir

$$g(x;\beta) = \underline{\underline{\frac{1}{\beta-\beta\exp\left\{-\frac{\pi}{2\beta}\right\}}\exp\left\{-\arctan(x)/\beta\right\}\cdot\frac{1}{1+x^2},\quad x>0.}}$$

Oppgave 3: Ultralydavbildning med kontrastmiddel

Vi har Y_i, $i = 1,\ldots,n$ u.i.f. normal$(y;\mu,\sigma)$.

a)

Her er $\mu = 1{,}0$ og $\sigma = 0{,}01$. Transformerer til standard normal vha $Z = (Y-\mu)/\sigma$.

$$P(Y_i > 1{,}0) = P\left(Z > \frac{1{,}0-1{,}0}{0{,}01}\right) = P(Z > 0{,}0) = \underline{\underline{0{,}5}}.$$

Dette kan ses fra tabellen, eller av symmetri i normalfordelingen, eller også direkte fra fordelingen til Y_i.

$$P(|Y_i - 1{,}0| > 0{,}02) = P(Y_i - 1{,}0 > 0{,}02 \cup Y_i - 1{,}0 < -0{,}02) = 2P(Y_i - 1{,}0 > 0{,}02).$$

Merk at $Y_i - 1{,}0$ tilsvarer $Y_i - \mu$, så divisjon med σ gir standard normalfordeling:

$$2P\left(\frac{Y_i - 1{,}0}{\sigma} > \frac{0{,}02}{\sigma}\right) = 2P(Z > 2) = 2(1 - P(Z \leq 2)) = \underline{\underline{0{,}046}}.$$

Gjennomsnittet $\bar{Y} = (Y_1 + Y_2)/2$ har samme forventningsverdi $\mu = 1{,}0$, men standardavvik $\sigma/\sqrt{2}$. Beregningen er ellers som over.

$$2P\left(\frac{\bar{Y} - 1{,}0}{\sigma/\sqrt{2}} > \frac{0{,}02}{\sigma/\sqrt{2}}\right) = 2P(Z > 2\sqrt{2}) \approx 2(1 - P(Z \leq 2{,}83)) = \underline{\underline{0{,}0046}}.$$

b)

Har nå $Y = \alpha + \beta x + E$ som i vanlig lineær regresjon. Dvs. $\mathrm{E}(Y_i) = \alpha + \beta x_i$ og $\mathrm{Var}(Y_i) = \sigma^2$.

Dette gir $\mathrm{E}(\bar{Y}) = \alpha + \beta\bar{x}$ og $\mathrm{Var}(\bar{Y}) = \sigma^2/n$.

$$\mathrm{E}(A) = \mathrm{E}(\bar{Y} - B\bar{x}) = \mathrm{E}(\bar{Y}) - \mathrm{E}(B)\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \underline{\underline{\alpha}}.$$

$$\mathrm{Var}(A) \overset{\text{uavh}}{=} \mathrm{Var}(\bar{Y}) + \mathrm{Var}(B)\bar{x}^2 = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \underline{\underline{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\ \sum_{i=1}^n (x_i - \bar{x})^2}}}.$$

Den siste overgangen kommer ved at når uttrykket settes på felles brøkstrek blir telleren $\sigma^2(n\bar{x}^2 + \sum_{i=1}^n (x_i - \bar{x})^2) = \sigma^2(n\bar{x}^2 + \sum_{i=1}^n x_i^2 - n\bar{x}^2)$. (Kvadrér ut $(x_i - \bar{x})^2$, del opp summen og bruk at $\sum_{i=1}^n x_i = n\bar{x}$.)

Fra definisjonen av kovarians får vi

$$\mathrm{Cov}(A, B) = \mathrm{E}\{(A - \mathrm{E}[A])(B - \mathrm{E}[B])\} = \mathrm{E}(AB) - \mathrm{E}(A)\,\mathrm{E}(B).$$

Merk forresten at $\mathrm{Cov}(B, B) = \mathrm{Var}(B)$.

Setter nå inn for A. B og $\bar{Y}$ er uavhengige, som medfører at $\mathrm{Cov}(\bar{Y}, B) = 0$.

$$\begin{aligned}
\mathrm{Cov}(\bar{Y} - B\bar{x}, B) &= \mathrm{E}((\bar{Y} - B\bar{x})B) - \mathrm{E}(\bar{Y} - B\bar{x})\,\mathrm{E}(B) \\
&= \mathrm{E}(\bar{Y}B) - \mathrm{E}(\bar{Y})\,\mathrm{E}(B) - \bar{x}[\mathrm{E}(B^2) - \mathrm{E}(B)^2] \\
&= \mathrm{Cov}(\bar{Y}, B) - \bar{x}\,\mathrm{Var}(B) = \underline{\underline{-\bar{x}\,\mathrm{Var}(B)}}.
\end{aligned}$$

Følgende linje vil også være fullgodt svar:

$$\mathrm{Cov}(A, B) = \mathrm{Cov}(\bar{Y} - B\bar{x}, B) = \mathrm{Cov}(\bar{Y}, B) - \bar{x}\,\mathrm{Cov}(B, B) = -\bar{x}\,\mathrm{Var}(B) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Kovariansen er negativ hvis $\bar{x}$ er positiv og vice versa.

Det er mulig å komme fram til rett svar ved å uttrykke B og $\bar{Y}$ med Y_i og løse ut slik at $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ og $\mathrm{Cov}(Y_i, Y_i) = \sigma^2$. Dette blir imidlertid mye mer regnekrevende.

c)

Bruk av formelsamlingen gir at

$$V = \frac{n\widetilde{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - A - Bx_i)^2$$

er χ^2-fordelt med $\nu = n-2$ frihetsgrader. Dette kan brukes for å finne forventningsverdien;

$$\mathrm{E}\left(\frac{n\widetilde{\sigma}^2}{\sigma^2}\right) = \mathrm{E}(V) = \nu = n-2 \Longrightarrow \mathrm{E}(\widetilde{\sigma}^2) = \underline{\underline{\frac{n-2}{n}\sigma^2}}.$$

Variansestimatoren er ikke lik den forventningsrette estimatoren $\widehat{\sigma}^2$ som vi kjenner fra forelesning, lærebok og formelsamlingen.

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - A - Bx_i)^2, \qquad \widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - A - Bx_i)^2.$$

Vi ser at $\widehat{\sigma^2} = \frac{n}{n-2}\widetilde{\sigma}^2$. Hvis du tar forventningsverdi på begge sider av dette uttrykket, og utnytter at vi vet $\mathrm{E}(\widehat{\sigma^2}) = \sigma^2$ blir $\mathrm{E}(\widetilde{\sigma}^2) = \frac{n-2}{n}\sigma^2$.

Vi har hypotesetesten

$$\begin{aligned} H_0 :& \quad \sigma^2 = \sigma_0^2 = 0{,}01^2, \\ H_1 :& \quad \sigma^2 > 0{,}01^2, \end{aligned}$$

Som testobservator bruker vi V som gitt over. Under H_0 er

$$V_0 = \frac{n\widetilde{\sigma}^2}{\sigma_0^2}$$

χ^2-fordelt med $\nu = n-2$ frihetsgrader.

Vi forkaster H_0 dersom observert verdi $v_0 > k$. Verdien k blir valgt slik at $P(V_0 > k \mid H_0) = \alpha$, som gir at $k = \chi^2_{n-2,\alpha}$. α er signifikansnivået.

Forkastningsområdet uttrykt for $\widetilde{\sigma}^2$ blir

$$\widetilde{\sigma}^2 > \frac{\sigma_0^2\chi^2_{28,0,01}}{n}.$$

Ved innsetting for v med $n = 30$, $\nu = 28$, $\alpha = 0{,}01$ og $\sigma^2 = \sigma_0^2 = 0{,}01^2$ blir forkastningsområdet

$$\widetilde{\sigma}^2 > \frac{\sigma_0^2 \chi^2_{28,0,01}}{n} = 0{,}000161 = 0{,}0127^2.$$

Dvs: dette er den verdien vi minst må observere for $\widetilde{\sigma}^2$ for at den antatt friske testpersonen likevel skal sendes til dyr kreftundersøkelse.

Kommentar: Oppgaven illustrerer at to gjennomsnitt av to målinger kan redusere halesannsynlighet svært mye, hvordan stigningstall og konstantledd er koplet sammen, og at vi godt kan regne konfidensintervall, hypotesetest osv. selv om estimatoren ikke er forventningsrett.