TMA4240 Statistikk
Høst 2020

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalte oppgaver 10
Løsningsskisse

Oppgave 1

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

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

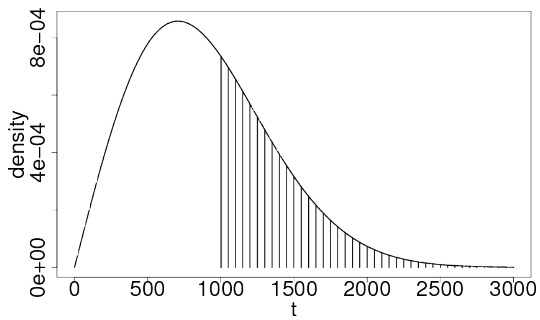
a)

$$\text{P}(T > 1000) = 1 - \text{P}(T \leq 1000) = 1 - F(1000) = 1 - \left(1 - \exp\left\{-\frac{2\cdot 1000^2}{2\cdot 10^6}\right\}\right) = e^{-1} = \underline{\underline{0.368}}$$

$$\text{P}(T > 2000|T > 1000) = \frac{\text{P}(T > 2000 \cap T > 1000)}{\text{P}(T > 1000)} = \frac{\text{P}(T > 2000)}{\text{P}(T > 1000)} = \frac{1 - F(2000)}{e^{-1}}$$
$$= \frac{1 - \left(1 - \exp\left\{-\frac{2\cdot 2000^2}{2\cdot 10^6}\right\}\right)}{e^{-1}} = \frac{e^{-4}}{e^{-1}} = e^{-4+1} = e^{-3} = \underline{\underline{0.050}}.$$

La Z være antall av de tre levetidene som er større enn 1000 døgn. Siden de tre levetidene er antatt uavhengige blir da Z binomisk fordelt med $n = 3$ forsøk og sannsynlighet for suksess lik $p = \text{P}(T > 1000) = e^{-1}$. Dermed får man at

$$\text{P}(Z \geq 2) = \text{P}(Z = 2) + \text{P}(Z = 3) = \binom{3}{2}p^2(1-p)^1 + \binom{3}{3}p^3(1-p)^0$$
$$= 3p^2(1-p) + p^3 = \underline{\underline{0.306}}.$$



Figur 1: Plott av $f(t)$ når $z = 2.0$ og $\theta = 2 \cdot 10^6$. Arealet av det skraverte området er lik sannsynligheten $\mathrm{P}(T > 1000)$.

b) For $t > 0$ får vi at

$$f(t) = F'(t) = -\exp\left\{-\frac{zt^2}{\theta}\right\} \cdot \left(-\frac{2zt}{\theta}\right) = \frac{2zt}{\theta}\exp\left\{-\frac{zt^2}{\theta}\right\}.$$

En skisse av sannsynligetstettheten $f(t)$ for $t \in [0, 3000]$ er gitt i figur 1. I denne figuren er arealet som er lik sannsynligheten $\mathrm{P}(T > 1000)$ skravert.

c) Den en-éntydige transformasjonen mellom t og v er gitt som

$$v = \frac{2zt^2}{\theta} \quad \Leftrightarrow \quad t = \sqrt{\frac{v\theta}{2z}}.$$

Merk at transformasjonen er en-éntydig fordi det er gitt at $t > 0$. For $v > 0$ gir transformasjonsformelen da at sannsynlighetstettheten for V blir

$$f_V(v) = f_T\left(\sqrt{\frac{v\theta}{2z}}\right) \cdot \left|\frac{\mathrm{d}t}{\mathrm{d}v}\right| = \frac{2z\sqrt{\frac{v\theta}{2z}}}{\theta}\exp\left\{-\frac{z\left(\sqrt{\frac{v\theta}{2z}}\right)^2}{\theta}\right\} \cdot \frac{1}{2\sqrt{\frac{v\theta}{2z}}}\frac{\theta}{2z} = \frac{1}{2}\exp\left\{-\frac{v}{2}\right\}.$$

Siden V ikke kan være negativ blir selvfølgelig $f_V(v) = 0$ for $v < 0$. Sannsynlighetstettheten til en χ^2-fordeling med ν frihetsgrader er gitt som

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2)}x^{\nu/2-1}e^{-x/2} \quad \text{for } x > 0.$$

Setter vi her inn $\nu = 2$ får vi at tettheten blir

$$f(x) = \frac{1}{2\Gamma(1)}x^{1-1}e^{-x/2} = \frac{1}{2}\exp\left\{-\frac{x}{2}\right\}.$$

Vi ser at dette er samme sannsynlighetstetthet som $f_V(v)$ som vi utledet over. Dermed har vi vist at $V \sim \chi^2_2$.

For en χ^2_v-fordeling vet vi generelt at forventingsverdien er lik ν og variansen er lik 2ν. Dermed har vi spesielt at $\mathrm{E}[V] = 2$ og $\mathrm{Var}[V] = 4$. Dette gir

$$\mathrm{E}[V] = \mathrm{E}\left[\frac{2zT^2}{\theta}\right] = \frac{2z}{\theta}\mathrm{E}[T^2] = 2 \quad \Rightarrow \quad \mathrm{E}[T^2] = \frac{\theta}{z}$$

og

$$\mathrm{Var}[V] = \mathrm{Var}\left[\frac{2zT^2}{\theta}\right] = \left(\frac{2z}{\theta}\right)^2 \mathrm{E}\left[T^2\right] = 4 \quad \Rightarrow \quad \mathrm{Var}\left[T^2\right] = \left(\frac{\theta}{z}\right)^2.$$

d) Rimelighetsfunksjonen blir gitt som

$$L(\theta) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n}\left[\frac{2z_i t_i}{\theta}\exp\left\{-\frac{z_i t_i^2}{\theta}\right\}\right].$$

Log-rimelighetsfunksjonen blir dermed

$$\begin{aligned} l(\theta) = \ln(L(\theta)) &= \sum_{i=1}^{n}\ln\left(\frac{2z_i t_i}{\theta}\exp\left\{-\frac{z_i t_i^2}{\theta}\right\}\right) \\ &= \sum_{i=1}^{n}\left[\ln(2) + \ln(z_i) + \ln(t_i) - \ln(\theta) - \frac{z_i t_i^2}{\theta}\right] \\ &= n\ln(2) + \sum_{i=1}^{n}\ln(z_i) + \sum_{i=1}^{n}\ln(t_i) - n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} z_i t_i^2. \end{aligned}$$

Finner maksimum ved å derivere med hensyn på θ og sette lik null,

$$l'(\theta) = -n\cdot\frac{1}{\theta} - \left(-\frac{1}{\theta^2}\right)\sum_{i=1}^{n} z_i t_i^2 = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} z_i t_i^2 = 0 \quad \Rightarrow \quad \theta = \frac{1}{n}\sum_{i=1}^{n} z_i t_i^2.$$

Sannsynlighetsmaksimeringsestimatoren for θ blir dermed

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} z_i T_i^2.$$

Benytter resultatene fra **c**) til å finne forventingsverdi og varians for $\widehat{\theta}$,

$$\begin{aligned} \mathrm{E}\left[\widehat{\theta}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} z_i T_i^2\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} z_i T_i^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[z_i T_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} z_i \mathrm{E}\left[T_i^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} z_i\frac{\theta}{z_i} = \frac{1}{n}\sum_{i=1}^{n}\theta = \frac{1}{n}\cdot n\theta = \theta \quad \Rightarrow \quad \widehat{\theta} \text{ er forventingsrett,} \end{aligned}$$

$$\mathrm{Var}\left[\widehat{\theta}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} z_i T_i^2\right] = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} z_i T_i^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[z_i T_i^2\right] = \frac{1}{n^2}\sum_{i=1}^{n} z_i^2 \mathrm{Var}\left[T_i^2\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} z_i^2 \text{Var}\left[T_i^2\right] = \frac{1}{n^2}\sum_{i=1}^{n} z_i^2 \left(\frac{\theta}{z_i}\right)^2 = \frac{\theta^2}{n^2}\sum_{i=1}^{n} 1 = \frac{\theta^2}{n^2}\cdot n = \underline{\underline{\frac{\theta^2}{n}}}.$$

Merk at man i utregningen av variansen har benyttet at T_i'ene er uavhengige.

e) Vi har at

$$U = \sum_{i=1}^{n} V_i \quad \text{der } V_i = \frac{2z_i T_i^2}{\theta}.$$

Fra punkt **c**) vet vi at $V_i \sim \chi_2^2$. Dessuten, siden T_i'ene er antatt uavhengige vil også V_i'ene være uavhengige. Siden en sum av uavhengige χ^2-fordelte variabler også blir χ^2-fordelt der antall frihetsgrader til summen er lik summen av antall frihetsgrader får vi dermed at U er χ^2-fordelt med $\sum_{i=1}^{n} 2 = 2n$ frihetsgrader.

Siden $U \sim \chi_{2n}^2$ får vi at

$$\text{P}\left(\chi^2_{1-\frac{\alpha}{2},2n} \leq U \leq \chi^2_{\frac{\alpha}{2},2n}\right) = 1-\alpha.$$

Setter vi inn uttrykket for U får vi

$$\text{P}\left(\chi^2_{1-\frac{\alpha}{2},2n} \leq \frac{2}{\theta}\sum_{i=1}^{n} z_i T_i^2 \leq \chi^2_{\frac{\alpha}{2},2n}\right) = 1-\alpha.$$

Løser de to ulikheten med hensyn på θ hver for seg,

$$\chi^2_{1-\frac{\alpha}{2},2n} \leq \frac{2}{\theta}\sum_{i=1}^{n} z_i T_i^2 \quad \Rightarrow \quad \theta \leq \frac{2}{\chi^2_{1-\frac{\alpha}{2},2n}}\sum_{i=1}^{n} z_i T_i^2,$$

$$\frac{2}{\theta}\sum_{i=1}^{n} z_i T_i^2 \leq \chi^2_{\frac{\alpha}{2},2n} \quad \Rightarrow \quad \frac{2}{\chi^2_{\frac{\alpha}{2},2n}}\sum_{i=1}^{n} z_i T_i^2 \leq \theta$$

Dermed har vi at

$$\text{P}\left(\frac{2}{\chi^2_{\frac{\alpha}{2},2n}}\sum_{i=1}^{n} z_i T_i^2 \leq \theta \leq \frac{2}{\chi^2_{1-\frac{\alpha}{2},2n}}\sum_{i=1}^{n} z_i T_i^2\right) = 1-\alpha$$

slik at et $(1-\alpha)\cdot 100\%$-konfidensintervall for θ er gitt ved

$$\underline{\underline{\left[\frac{2}{\chi^2_{\frac{\alpha}{2},2n}}\sum_{i=1}^{n} z_i T_i^2, \frac{2}{\chi^2_{1-\frac{\alpha}{2},2n}}\sum_{i=1}^{n} z_i T_i^2\right]}}.$$

For $\alpha = 0.05$ og $n = 10$ finner vi i tabell at $\chi^2_{1-\frac{\alpha}{2},2n} = \chi^2_{0.975,20} = 9.591$ og $\chi^2_{\frac{\alpha}{2},2n} = \chi^2_{0.025,20} = 34.170$. Innsatt oppgitte observasjoner blir dermed konfidensintervallet

$$\left[\frac{2}{34.170}\cdot 23\ 287\ 125, \frac{2}{9.591}\cdot 23\ 287\ 125\right] = \underline{\underline{[1\ 363\ 016, 4\ 856\ 037]}}.$$

f) Ved å ta utgangspunkt i Y har vi at

$$\mathrm{P}\left(y_{1-\frac{\alpha}{2}} \le Y \le y_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Setter vi inn uttrykket for Y, etter å ha forenklet dette ved å forkorte, får vi dermed

$$\mathrm{P}\left(y_{1-\frac{\alpha}{2}} \le \frac{nz_0T_0^2}{\sum_{i=1}^n z_iT_i^2} \le y_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Løser de to ulikhetene med hensyn på T_0 hver for seg (og husker på at vi vet at $T_0 > 0$),

$$y_{1-\frac{\alpha}{2}} \le \frac{nz_0T_0^2}{\sum_{i=1}^n z_iT_i^2} \quad \Rightarrow \quad \sqrt{\frac{y_{1-\frac{\alpha}{2}}\sum_{i=1}^n z_iT_i^2}{nz_0}} \le T_0,$$

$$\frac{nz_0T_0^2}{\sum_{i=1}^n z_iT_i^2} \le y_{\frac{\alpha}{2}} \quad \Rightarrow \quad T_0 \le \sqrt{\frac{y_{\frac{\alpha}{2}}\sum_{i=1}^n z_iT_i^2}{nz_0}}.$$

Dermed har vi at

$$\mathrm{P}\left(\sqrt{\frac{y_{1-\frac{\alpha}{2}}\sum_{i=1}^n z_iT_i^2}{nz_0}} \le T_0 \le \sqrt{\frac{y_{\frac{\alpha}{2}}\sum_{i=1}^n z_iT_i^2}{nz_0}}\right) = 1-\alpha,$$

slik at et $(1-\alpha)\cdot 100\%$-prediksjonsintervall for T_0 er

$$\underline{\underline{\left[\sqrt{\frac{y_{1-\frac{\alpha}{2}}\sum_{i=1}^n z_iT_i^2}{nz_0}}, \sqrt{\frac{y_{\frac{\alpha}{2}}\sum_{i=1}^n z_iT_i^2}{nz_0}}\right]}}.$$

For å få et 90%-prediksjonsintervall må vi velge $\alpha = 0.1$ og da blir $y_{1-\frac{\alpha}{2}} = y_{0.95} = 0.051$ og $y_{\frac{\alpha}{2}} = y_{0.05} = 3.49$. Innsatt oppgitte observasjoner blir dermed prediksjonsintervallet

$$\left[\sqrt{\frac{0.051\cdot 23\ 287\ 125}{10\cdot 3}}, \sqrt{\frac{3.49\cdot 23\ 287\ 125}{10\cdot 3}}\right] = \underline{\underline{[198.97, 1\ 645.92]}}.$$