TMA4245 Statistikk
Eksamen 21. mai 2013

Korrigert 10. juni 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Løsningsskisse

Oppgave 1

Et plott av sannsynlighetstetthene er gitt i figur 1. Videre har vi

$$\mathrm{P}(X \leq 1.2) = \Phi(1.2) = \underline{\underline{0.8849}}$$

og

$$\mathrm{P}(Y > 2) = 1 - \mathrm{P}(Y \leq 2) = 1 - \Phi\left(\frac{2-1}{2}\right) = 1 - \Phi\left(\frac{1}{2}\right) = 1 - 0.6915 = \underline{\underline{0.3085}}.$$

Siden X og Y er antatt uavhengige og normalfordelte, og $X+Y$ er en lineærkombinasjon av X og Y, får vi at $X+Y$ også er normalfordelt. Dessuten får vi at $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 0 + 1 = 1$ og siden X og Y er uavhengige, $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] = 1^2 + 2^2 = 5$. Dermed blir

$$\mathrm{P}(X+Y \leq 2) = \Phi\left(\frac{2-1}{\sqrt{5}}\right) = \Phi(0.45) = \underline{\underline{0.6736}}.$$

Oppgave 2

Multiplikasjonssetningen sier at

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).$$

Dermed får vi at

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cup B) = 0.2 + 0.5 - 0.6 = 0.1 \neq 0.$$

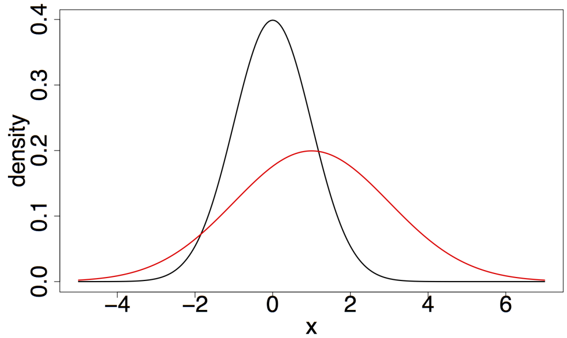
Siden $\mathrm{P}(A \cap B) \neq 0$ er hendelsene A og B ikke disjunkte.

Hendelsene A og B er uavhengige hvis og bare hvis $\mathrm{P}(A|B)$ er lik $\mathrm{P}(A)$. Må derfor regne ut $\mathrm{P}(A|B)$,

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{0.1}{0.5} = \frac{1}{5} = 0.2 = \mathrm{P}(A).$$

Hendelsene A og B er dermed uavhengige.

Oppgave 3



Figur 1: Svart kurve er sannsynlighetstettheten for en normalfordeling med forventningsverdi lik 0 og standardavvik lik 1. Rød kurve er sannsynlighetstettheten for en normalfordeling med forventningsverdi lik 1 og standardavvik lik 2.

a)

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

$$\mathrm{P}(T > 2000 | T > 1000) = \frac{\mathrm{P}(T > 2000 \cap T > 1000)}{\mathrm{P}(T > 1000)} = \frac{\mathrm{P}(T > 2000)}{\mathrm{P}(T > 1000)} = \frac{1 - F(2000)}{e^{-1}}$$

$$= \frac{1 - \left(1 - \exp\left\{-\frac{2 \cdot 2000^2}{2 \cdot 10^6}\right\}\right)}{e^{-1}} = \frac{e^{-4}}{e^{-1}} = e^{-4+1} = e^{-3} = \underline{\underline{0.050}}.$$

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 $p = \mathrm{P}(T > 1000) = e^{-1}$. Dermed får man at

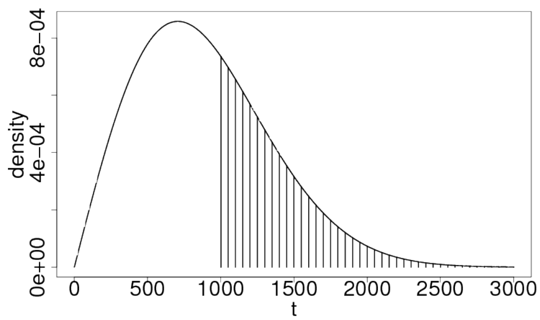
$$\mathrm{P}(Z \ge 2) = \mathrm{P}(Z = 2) + \mathrm{P}(Z = 3) = \binom{3}{2} p^2(1-p)^1 + \binom{3}{3} p^3 (1-p)^0$$

$$= 3p^2(1-p) + p^3 = \underline{\underline{0.306}}.$$

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 2. I denne figuren er arealet som er lik sannsynligheten $\mathrm{P}(T > 1000)$ skravert.



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

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 $\widehat{\theta}$,

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

$$\begin{aligned} \mathrm{Var}\left[\widehat{\theta}\right] &= \mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^{n} z_i T_i^2 \right] = \left(\frac{1}{n} \right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} z_i T_i^2 \right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left[z_i T_i^2 \right] = \frac{1}{n^2} \sum_{i=1}^{n} z_i^2 \mathrm{Var}\left[T_i^2 \right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} z_i^2 \mathrm{Var}\left[T_i^2 \right] = \frac{1}{n^2} \sum_{i=1}^{n} z_i^2 \left(\frac{\theta}{z_i} \right)^2 = \frac{\theta^2}{n^2} \sum_{i=1}^{n} 1 = \frac{\theta^2}{n^2} \cdot n = \underline{\underline{\frac{\theta^2}{n}}}. \end{aligned}$$

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 $U \sim \chi^2_{2n}$ får vi at

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

Setter vi inn uttrykket for U får vi

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

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

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

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

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

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

$$\mathrm{P}\left(y_{1-\frac{\alpha}{2}} \leq \frac{n z_0 T_0^2}{\sum_{i=1}^n z_i T_i^2} \leq y_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

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

$$y_{1-\frac{\alpha}{2}} \leq \frac{nz_0 T_0^2}{\sum_{i=1}^n z_i T_i^2} \quad \Rightarrow \quad \sqrt{\frac{y_{1-\frac{\alpha}{2}} \sum_{i=1}^n z_i T_i^2}{nz_0}} \leq T_0,$$

$$\frac{nz_0 T_0^2}{\sum_{i=1}^n z_i T_i^2} \leq y_{\frac{\alpha}{2}} \quad \Rightarrow \quad T_0 \leq \sqrt{\frac{y_{\frac{\alpha}{2}} \sum_{i=1}^n z_i T_i^2}{nz_0}}.$$

Dermed har vi at

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

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

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

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

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

Oppgave 4

a) Siden man skal undersøke om "de observerte data gir grunnlag for å påstå at forventet løpetid **avtar** med økende antall armhevinger" må man velge som $\beta_1 < 0$ som H_1. Vi har dermed

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 < 0.$$

Bruker testobservator

$$T = \frac{\widehat{\beta}_1}{\sqrt{\frac{S^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

og vi vet at denne er t-fordelt med $n-2$ frihetsgrader når H_0 er riktig.

Forkaster H_0 dersom $T < k$ der kritisk verdi k bestemmes fra kravet

$$\mathrm{P}(\text{Forkast } H_0 | H_0 \text{ er riktig}) = \alpha = 0.05,$$

$$\mathrm{P}(T < k | H_0 \text{ er riktig}) = 0.05.$$

Vi må dermed ha $k = t_{1-0.05,n-2} = -t_{0.05,n-2}$. $\underline{\underline{\text{Vi skal dermed forkaste } H_0 \text{ dersom } T < -t_{0.05,n-2}}}$.

Med våre data har vi $n = 42$,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{-12163.6}{11113.9} = -1.094$$

og

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = \frac{414563.1}{40} = 10364.1.$$

Observert verdi for testobservatoren blir dermed

$$t = \frac{-1.094}{\sqrt{\frac{10361.1}{11113.9}}} = -1.13.$$

Kritisk verdi når $n = 42$ finner vi i tabellen over t-fordelingen til å være $t_{0.05,n-2} = t_{0.05,40} = 1.684$. Siden $t = -1.13 \not< -1.684$ skal man ikke forkaste H_0. Dvs.

ikke grunnlag for å påstå at forventet løpetid avtar med økende antall armhevinger.

b) Ut fra begge plottene i figur 2 ser vi at fordelingen til ε_i ikke synes å være symmetrisk fordelt omkring null slik modellen antar. Mer spesifikt ser vi at fordelingen til residualene synes å ha en tyngre hale mot høyre enn mot venstre. Antagelsen om at alle residualene har lik fordeling synes tilfredsstilt.