TMA4245 Statistikk
Vår 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 9, blokk II
Løsningsskisse

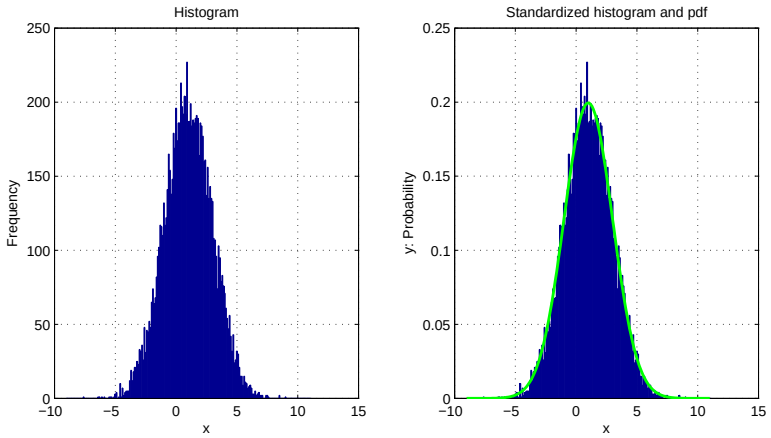
Oppgave 1

Scriptet *run_confds.m* simulerer n data $x_1, \ldots, x_n$ fra en normalfordeling med forventningsverdi $\mu = 1$ og varians $\sigma^2 = 2^2$ ved å trekke n ganger fra en standard normalfordeling $y_i \sim N(0,1)$ og utføre lineærtransformasjonen

$$x_i = \mu + \sigma \cdot y_i \ , \ \ i = 1, \ldots, n$$

Fra uttrykket kan vi greit regne på at da vil $x_i \sim N(\mu, \sigma^2)$. *(I Matlab trekker man fra en standard normalfordeling med funksjonen 'randn').*

Kjører vi scriptet får vi et histogram av $n = 10000$ simulerte data $x_1, \ldots, x_n$, som f.eks. kan se slik ut

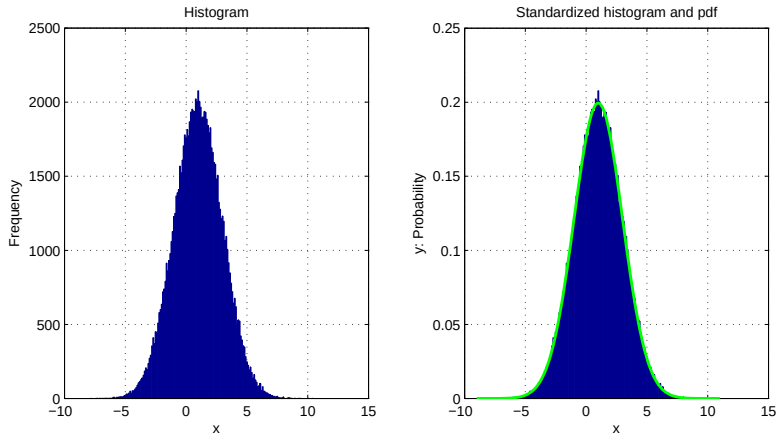


Figur 1: Histogram av $n = 10000$ simulerte data fra $N(1, 2^2)$

Histogrammet til høyre er standardisert, altså transformert slik at areal under histogramsøylene blir 1. I plottet er det i grønt også tegnet inn kurven for normalfordelingen med forventning 1 og standardavvik 2. Vi ser at de simulerte dataene overlapper normalfordelingen de kommer fra veldig bra. Dette siden vi simulerer såpass mange datapunkter. Det resulterende gjennomsnittet $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1.0047$ er veldig nærme den sanne forventnings-

verdien som også ligger innenfor det estimerte konfidensintervalet [0.96591.0434].

Trekker vi stedet $n = 100000$ data (setter altså parameteren 'n' i scriptet til 100000) kan histogrammet f.eks. se ut som i Fig.2 med estimert forventningsverdi $\hat{\mu} = 0.9983$ og estimert 95% konfidensinterval $[0.9859, 1.0107]$. Igjen er estimatet tilnærmet likt sann forventnigsverdi, som ligger innenfor konfidensintervalet, og overlappen mellom dataene og normalkurven er enda bedre.



Figur 2: Histogram av $n = 100000$ simulerte data fra $N(1, 2^2)$

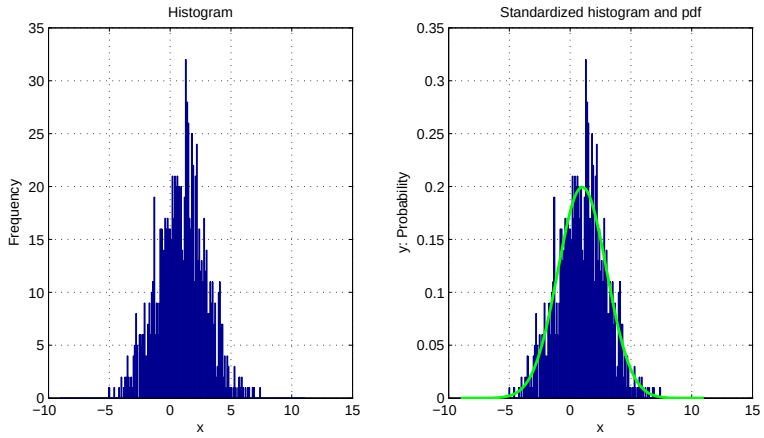
Trekker vi $n = 1000$ data (setter altså parameteren 'n' i scriptet til 1000) kan histogrammet f.eks. se ut som i Fig.3. med estimert forventningsverdi $\hat{\mu} = 0.9594$ og estimert 95% konfidensinterval [0.83741.0815]. Estimatet er fortsatt bra, men ikke like nærme som i tilfellene med høyere n. Vi ser også at estimert konfidensinterval er litt bredere, og at overlappen mellom dataene og normalkurven er dårligere (dette er også fordi vi har så liten oppløsning på histogrammet).

Det estimerte konfidensintervalet er beregnet som

$$\left[\hat{\mu} - 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}} \ , \ \hat{\mu} + 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

Når datamengden vokser og estimatet på standardaviket ikke varierer mye ser vi at faktoren $\frac{\hat{\sigma}}{\sqrt{n}}$ går mot 0, altså blir konfidensintervalet smalere jo større datamengden er. Vi merker oss også at vi her har brukt kvantilen $z_{0.025} = 1.96$ fra en normalfordeling selv om vi her bruker estimert varians. Med ukjent varians burde vi egentlig brukt kvantiler fra t-fordeling, men siden datamangden er så stor ($n \geq 1000$) vil t-fordeling med $n - 1$ frihetsgrader være tilnærmet lik standard normalfordeling.

Oppgave 2



Figur 3: Histogram av $n = 1000$ simulerte data fra $N(1, 2^2)$

a)

$$\begin{aligned} P(X < 6.74) &= P(\frac{X - 6.8}{0.06} < \frac{6.74 - 6.8}{0.06}) \\ &= \Phi(-1) = 1 - \Phi(1) \\ &= 1 - 0.841 = 0.159 \end{aligned}$$

$$\begin{aligned} P(6.74 < X < 6.86) &= P(X < 6.86) - P(X < 6.74) \\ &= P(\frac{X - 6.8}{0.06} < \frac{6.86 - 6.8}{0.06}) - 0.159 \\ &= \Phi(1) - 0.159 = 0.841 - 0.159 = 0.682 \end{aligned}$$

$$\begin{aligned} P(|X - \mu|) > 0.06) &= P(X - \mu < -0.06) + P(X - \mu > 0.06) \\ &= P(\frac{X - \mu}{0.06} < -1) + P(\frac{X - \mu}{0.06} > 1) \\ &= \Phi(-1) + 1 - \Phi(1) = 2(1 - \Phi(1)) = 0.318 \end{aligned}$$

Eventuelt

$$\begin{aligned} P(|X - \mu|) > 0.06) &= 1 - P(6.74 < X < 6.86) \\ &= 1 - 0.682 = 0.318 \end{aligned}$$

b) $Y \sim N(\mu, \frac{\sigma^2}{5})$

$$\begin{aligned} P(|Y - \mu|) > 0.06) &= 2P(Y - \mu > 0.06) \\ &= 2(1 - P(\frac{Y - \mu}{\frac{0.06}{\sqrt{5}}} \leq \sqrt(5))) \\ &= 0.026 \end{aligned}$$

$Y = \frac{1}{5}\sum_{i=1}^{5} X_i$ er lineærkombinasjon av uavhengige normalfordelte variable. Dermed er Y normalfordelt med $E(Y) = \mu$ og $Var(Y) = \frac{\sigma^2}{5}$
$\Longrightarrow \frac{Y-\mu}{\frac{\sigma}{\sqrt{5}}} \sim N(0,1)$
$\Longrightarrow P(-Z_{0.025} < \frac{y-\mu}{\frac{\sigma}{\sqrt{5}}} < Z_{0.025}) = 0.95$

$P(Y - Z_{0.025}\frac{\sigma}{\sqrt{5}} < \mu < Y + Z_{0.025}\frac{\sigma}{\sqrt{5}}) = 0.95$
D.v.s 95% konf. int. blir:

$$[Y - Z_{0.025}\cdot\frac{\sigma}{\sqrt{5}}, Y + Z_{0.025}\frac{\sigma}{\sqrt{5}}]$$

Innsatt tall:
$y = \bar{x} = 6.76, \sigma = 0.06, z_{0.025} = 1.96$

$[6.76 - (1.96)\frac{0.06}{\sqrt{5}}, 6.76 + (1.96)\frac{0.06}{\sqrt{5}}] = [6.707, 6.813]$

Oppgave 3

a) La V være målt vekt, slik at $V \sim N(\mu, \sigma^2) = N(10, 0.2^2)$. Vi får

$$\begin{aligned} P(V > 10.2) &= P\left(\frac{V-\mu}{\sigma} > \frac{10.2-10}{0.2}\right) = P(Z > 1) \\ &= 1 - P(Z \leq 1) = 1 - 0.8413 = \underline{\underline{0.1587}} \end{aligned}$$

$$\begin{aligned} P(|V-\mu| > 0.2) &= P(V - \mu > 0.2) + P(V - \mu < -0.2) \\ &= P\left(\frac{V-\mu}{\sigma} > \frac{0.2}{0.2}\right) + P\left(\frac{V-\mu}{\sigma} < -\frac{0.2}{0.2}\right) \\ &= P(Z > 1) + P(Z \leq -1) = 1 - P(Z \leq 1) + P(Z \leq -1) \\ &= 2 \cdot P(Z \leq -1) = 2 \cdot 0.1587 = \underline{\underline{0.3174}} \end{aligned}$$

La $\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$, slik at $\bar{V} \sim N(\mu, \sigma^2/n)$. Vi får

$$\begin{aligned} P(|\bar{V}-\mu| > 0.2) &= P(\bar{V} - \mu > 0.2) + P(\bar{V} - \mu < -0.2) \\ &= P\left(\frac{\bar{V}-\mu}{\sigma/\sqrt{n}} > \frac{0.2}{0.2/\sqrt{2}}\right) + P\left(\frac{\bar{V}-\mu}{\sigma/\sqrt{n}} < -\frac{0.2}{0.2/\sqrt{2}}\right) \\ &= P(Z > \sqrt{2}) + P(Z \leq -\sqrt{2}) \\ &= 1 - P(Z \leq \sqrt{2}) + P(Z \leq -\sqrt{2}) \\ &= 2 \cdot P(Z \leq -1.41) = 2 \cdot 0.0793 = \underline{\underline{0.1586}} \end{aligned}$$

b) Vi har $X_1 \sim N(\mu_A, \sigma^2)$ og $X_2 \sim N(\mu_B, \sigma^2)$ som er uavhengig av hverandre. Vi får ved fremgangsmåte 1:

$$\begin{aligned} \mathrm{E}[\hat{\mu}_A] &= \mathrm{E}[X_1] = \mu_A \\ \mathrm{Var}[\hat{\mu}_A] &= \mathrm{Var}[X_1] = \sigma^2 \end{aligned}$$

$$\mathrm{E}[\hat{\mu}_B] = \mathrm{E}[X_2] = \mu_B$$
$$\mathrm{Var}[\hat{\mu}_B] = \mathrm{Var}[X_2] = \sigma^2$$

Vi har $Y_1 \sim N(\mu_A + \mu_B, \sigma^2)$ og $Y_2 \sim N(\mu_A - \mu_B, \sigma^2)$ som er uavhengig av hverandre. Vi får ved fremgangsmåte 2:

$$\mathrm{E}[\tilde{\mu}_A] = \mathrm{E}[(Y_1 + Y_2)/2] = \frac{1}{2}(\mathrm{E}[Y_1] + \mathrm{E}[Y_2]) = \frac{1}{2}(\mu_A + \mu_B + \mu_A - \mu_B) = \mu_A$$

$$\mathrm{Var}[\tilde{\mu}_A] = \mathrm{Var}[(Y_1 + Y_2)/2] = \frac{1}{4}(\mathrm{Var}[Y_1] + \mathrm{Var}[Y_2]) = \frac{1}{4}(\sigma^2 + \sigma^2) = \sigma^2/2$$

$$\mathrm{E}[\tilde{\mu}_B] = \mathrm{E}[(Y_1 - Y_2)/2] = \frac{1}{2}(\mathrm{E}[Y_1] - \mathrm{E}[Y_2]) = \frac{1}{2}(\mu_A + \mu_B - \mu_A + \mu_B) = \mu_B$$

$$\mathrm{Var}[\tilde{\mu}_B] = \mathrm{Var}[(Y_1 - Y_2)/2] = \frac{1}{4}(\mathrm{Var}[Y_1] + \mathrm{Var}[Y_2]) = \frac{1}{4}(\sigma^2 + \sigma^2) = \sigma^2/2$$

Begge fremgangsmåtene gir forventningsrette estimatorer, så vi velger den med minst varians, dvs. fremgangsmåte 2: $\tilde{\mu}_A$ og $\tilde{\mu}_B$.

c) Vi har $\tilde{\mu}_A = u_1(Y_1, Y_2) = (Y_1 + Y_2)/2$ og $\tilde{\mu}_B = u_2(Y_1, Y_2) = (Y_1 - Y_2)/2$, som gir oss at $Y_1 = w_1(\tilde{\mu}_A, \tilde{\mu}_B) = \tilde{\mu}_A + \tilde{\mu}_B$ og $Y_2 = w_2(\tilde{\mu}_A, \tilde{\mu}_B) = \tilde{\mu}_A - \tilde{\mu}_B$. Fra transformasjonsformelen for to variabler har vi da at

$$g_{\tilde{\mu}_A, \tilde{\mu}_B}(\tilde{\mu}_A, \tilde{\mu}_B) = f_{Y_1, Y_2}(w_1(\tilde{\mu}_A, \tilde{\mu}_B), w_2(\tilde{\mu}_A, \tilde{\mu}_B)) \cdot |J|$$

hvor

$$J = \begin{vmatrix} \delta w_1/\delta\tilde{\mu}_A & \delta w_1/\delta\tilde{\mu}_B \\ \delta w_2/\delta\tilde{\mu}_A & \delta w_2/\delta\tilde{\mu}_B \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2.$$

Siden Y_1 og Y_2 er uavhengige, har vi $f_{Y_1,Y_2}(y_1,y_1) = f_{Y_1}(y_1)f_{Y_2}(y_2)$ og vi får følgende:

$$\begin{aligned}
g_{\tilde{\mu}_A,\tilde{\mu}_B}(\tilde{\mu}_A,\tilde{\mu}_B) &= f_{Y_1,Y_2}(w_1(\tilde{\mu}_A,\tilde{\mu}_B), w_2(\tilde{\mu}_A,\tilde{\mu}_B)) \cdot |J| \\
&= f_{Y_1}(w_1(\tilde{\mu}_A,\tilde{\mu}_B)) f_{Y_2}(w_2(\tilde{\mu}_A,\tilde{\mu}_B)) \cdot |-2| \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\tilde{\mu}_A+\tilde{\mu}_B-(\mu_A+\mu_B))^2\right\} \\
&\quad \cdot \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\tilde{\mu}_A-\tilde{\mu}_B-(\mu_A-\mu_B))^2\right\}\cdot 2 \\
&= \left(\frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sigma}\right)^2 \exp\left\{-\frac{1}{2\sigma^2}\left[(\tilde{\mu}_A+\tilde{\mu}_B)^2 - 2((\tilde{\mu}_A+\tilde{\mu}_B)\right.\right. \\
&\quad \cdot(\mu_A+\mu_B) + (\mu_A+\mu_B)^2 + (\tilde{\mu}_A-\tilde{\mu}_B)^2 - 2(\tilde{\mu}_A-\tilde{\mu}_B) \\
&\quad \left.\left.\cdot(\mu_A-\mu_B) + (\mu_A-\mu_B)^2\right]\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sigma}\right)^2 \exp\left\{-\frac{1}{2\sigma^2}\left[\tilde{\mu}_A^2 + 2\tilde{\mu}_A\tilde{\mu}_B + \tilde{\mu}_B^2 - 2\tilde{\mu}_A\mu_A\right.\right. \\
&\quad -2\tilde{\mu}_A\mu_B - 2\tilde{\mu}_B\mu_B + \mu_A^2 + 2\mu_A\mu_B + \mu_B^2 + \tilde{\mu}_A^2 - 2\tilde{\mu}_A\tilde{\mu}_B \\
&\quad +\tilde{\mu}_B^2 - 2\tilde{\mu}_A\mu_A + 2\tilde{\mu}_A\mu_B + 2\tilde{\mu}_B\mu_A - 2\tilde{\mu}_B\mu_B + \tilde{\mu}_A^2 \\
&\quad \left.\left.-2\mu_A\mu_B + \mu_B^2\right]\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sigma}\right)^2 \exp\left\{-\frac{1}{2\sigma^2}\left[2\tilde{\mu}_A^2 + 2\tilde{\mu}_B^2 - 4\tilde{\mu}_A\mu_A - 4\tilde{\mu}_B\mu_B\right.\right. \\
&\quad \left.\left.+2\tilde{\mu}_A^2 + 2\tilde{\mu}_B^2\right]\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sigma}\right)^2 \exp\left\{-\frac{2}{2\sigma^2}\left[(\tilde{\mu}_A-\mu_A)^2 + (\tilde{\mu}_B-\mu_B)^2\right]\right\} \\
&= \frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sigma}\exp\left\{-\frac{2}{2\sigma^2}(\tilde{\mu}_A-\mu_A)^2\right\} \\
&\quad \cdot \frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sigma}\exp\left\{-\frac{2}{2\sigma^2}(\tilde{\mu}_B-\mu_B)^2\right\} \\
&= g_{\tilde{\mu}_A}(\tilde{\mu}_A) g_{\tilde{\mu}_B}(\tilde{\mu}_B)
\end{aligned}$$

og dermed er $\tilde{\mu}_A$ og $\tilde{\mu}_B$ uavhengige ($\tilde{\mu}_A \sim N(\mu_A, \sigma^2/2)$ og $\tilde{\mu}_B \sim N(\mu_B, \sigma^2/2)$).

Oppgave 4

a) La $Z = 2\lambda T = u(T)$, som er en strengt monoton og deriverbar funksjon for alle T. Vi har $T = Z/(2\lambda) = w(Z)$ og $w'(Z) = 1/(2\lambda)$. Dette gir

$$g_Z(z) = f(w(z))|w'(z)| = \lambda e^{-\lambda(z/(2\lambda))}(1/(2\lambda)) = \begin{cases} \frac{1}{2}e^{-z/2} & , z > 0 \\ 0 & , \text{ellers} \end{cases}$$

b) Vi har $2\lambda T \sim \chi^2_2$. Dersom levetiden til komponontentene T_i er uavhengig, kan vi bruke

følgende resultat

$$\sum_{i=1}^{n} 2\lambda T_i \sim \chi^2_{\sum_{i=1}^{n} 2} = 2\lambda \sum_{i=1}^{n} T_i \sim \chi^2_{2n}$$

Vi finner et $1-\alpha$ konfidensintervall fra

$$\begin{aligned}
&P\left(\chi^2_{1-\alpha/2,2n} < 2\lambda \sum_{i=1}^{n} t_i < \chi^2_{\alpha/2,2n}\right) = 1-\alpha \\
&P\left(\frac{\chi^2_{1-\alpha/2,2n}}{2\sum_{i=1}^{n} t_i} < \lambda < \frac{\chi^2_{\alpha/2,2n}}{2\sum_{i=1}^{n} t_i}\right) = 1-\alpha \\
&P\left(\frac{\chi^2_{0.95/2,20}}{2 \cdot 6430.2} < \lambda < \frac{\chi^2_{0.05,20}}{2 \cdot 2430.2}\right) = 0.90 \\
&P\left(\frac{10.851}{12860.4} < \lambda < \frac{31.410}{12860.4}\right) = 0.90 \\
&P\left(8.438 \cdot 10^{-4} < \lambda < 24.424 \cdot 10^{-4}\right) = 0.90
\end{aligned}$$

Så et 90% konfidensintervall for λ er $\underline{\underline{(8.438 \cdot 10^{-4}, 24.424 \cdot 10^{-4})}}$.