



Norges teknisk-naturvitenskapelige universitet  
Institutt for matematiske fag

TMA4240 Statistikk  
Høst 2017

Anbefalt øving 9  
Løsningskisse

### Oppgave 1

- a) Vi lar her  $Y$  være antall fugler som kolliderer med vindmølla i løpet av den gitte perioden på  $t = 5$  uker. Siden  $Y$  er poissonfordelt med intensitet  $\lambda = 1$  fugl/uke, har  $Y$  punktsannsynlighet

$$P(Y = y) = \frac{(\lambda t)^y}{y!} e^{-\lambda t} = \frac{5^y}{y!} e^{-5}, \quad y = 0, 1, 2, \dots$$

Figur 1 illustrerer denne punktsannsynligheten for verdier av  $y$  mellom 0 og 14. Sannsynligheten for at mer enn 10 fugler kolliderer i løpet av vedkommende periode er da

$$P(Y > 10) = 1 - P(Y \leq 10) = 1 - \sum_{y=0}^{10} \frac{5^y}{y!} e^{-5} = 1 - 0.986 = \underline{\underline{0.014}}.$$

Sannsynligheten for at færre enn fem fugler kolliderer er

$$P(Y < 5) = P(Y \leq 4) = 0.44,$$

og den betingede sannsynligheten for at ingen fugler kolliderer, gitt at færre enn fem gjør det, er

$$P(Y = 0 | Y < 5) = \frac{P(Y = 0 \cap Y < 5)}{P(Y < 5)} = \frac{P(Y = 0)}{P(Y < 5)} = \frac{e^{-5}}{0.44} = \underline{\underline{0.015}}.$$

- b) Rimelighetsfunksjonen er sannsynligheten for å få  $Y = 261$  med parameter  $4 \cdot 52\lambda = 208\lambda$ . Rimelighetsfunksjonen er en funksjon av parameteren  $\lambda$ .

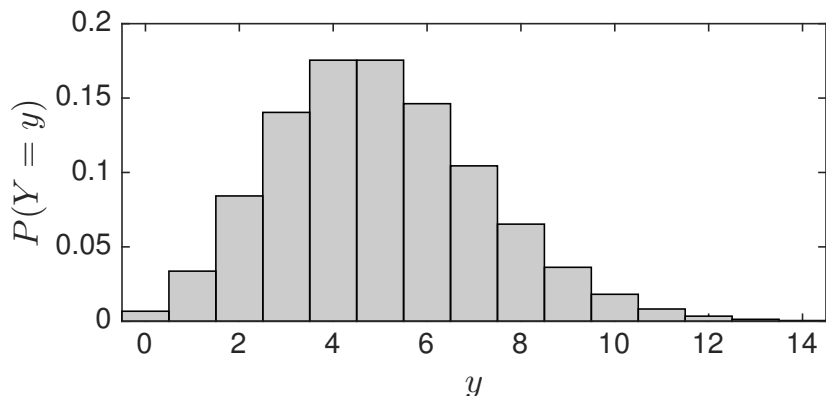
Vi tar logaritmen til rimelighetsfunksjonen for å forenkle utregningene,

$$l(\lambda) = \log P(Y = 261; \lambda) = 261 \log(208\lambda) - \log(261!) - 208\lambda$$

Vi finner maksimum ved å derivere med hensyn på  $\lambda$ ,

$$l'(\lambda) = \frac{261}{\lambda} - 208 = 0 \Rightarrow \lambda = \frac{261}{208}.$$

Dette gir  $\hat{\lambda} = 261/208 = 1.25$ .



Figur 1: Punktsannsynligheten til den stokastiske variabelen  $Y$ , som er poissonfordelt med forventningsverdi  $\mu = \lambda \cdot t = 5$ .

- c) Momentgenererende funksjon til en sum av to uavhengige variabler er produktet av de momentgenererende funksjonene for de to variablene. Momengenererende funksjon for en Poisson-fordelt variabel er

$$M_X(t) = \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y}{y!} e^{-\lambda} = \sum_{y=0}^{\infty} \frac{(e^t \lambda)^y}{y!} e^{-\lambda} = e^{\lambda(e^t-1)} \sum_{y=0}^{\infty} \frac{(e^t \lambda)^y}{y!} e^{-e^t \lambda} = e^{\lambda(e^t-1)}$$

hvor den siste summen er 1 fordi det er en sum over en Poisson-fordelt variabel med parameter  $e^t \lambda$ . Da er

$$M_Z(t) = M_X(t)M_Y(t) = e^{\lambda(e^t-1)} e^{\nu(e^t-1)} = e^{(\lambda+\nu)(e^t-1)}$$

Vi kjenner igjen formen på funksjonen som en momentgenererende funksjon. Dette er momentgenererende funksjon til en Poisson-fordelt tilfeldig variabel med parameter  $\lambda + \nu$ .

## Oppgave 2

- a) Sannsynligheten for at ventetiden er lenger enn 2 uker:

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - (1 - \exp(-0.04 \cdot 2^2)) = \exp(-0.16) = \underline{\underline{0.8521}}.$$

Sannsynligheten for at du må vente minst 5 uker, gitt at du må vente i minst 2 uker:

$$\begin{aligned} P(X > 5 | X > 2) &= \frac{P(X > 5 \cap X > 2)}{P(X > 2)} = \frac{P(X > 5)}{P(X > 2)} = \frac{1 - P(X \leq 5)}{P(X > 2)} \\ &= \frac{1 - F(5)}{P(X > 2)} = \frac{1 - (1 - \exp(-0.04 \cdot 5^2))}{0.8521} = \frac{0.3679}{0.8521} = \underline{\underline{0.4317}}. \end{aligned}$$

Sannsynlighetstettheten til  $X$  for  $x \geq 0$  finner vi ved å derivere  $F(x)$ :

$$f(x) = \frac{dF(x)}{dx} = 0 - (-2\alpha x \exp(-\alpha x^2)) = 2\alpha x \exp(-\alpha x^2), \quad \text{for } x \geq 0.$$

b) SME for  $\alpha$ :

Finner først rimelighetsfunksjonen, som er simultanfordelingen for  $X_1, X_2, \dots, X_n$  sett på som funksjon av  $x_i$ 'ene og  $\alpha$ :

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \alpha) &= f(x_1, x_2, \dots, x_n; \alpha) = \prod_{i=1}^n f(x_i; \alpha) \\ &= \prod_{i=1}^n 2\alpha x_i \exp(-\alpha x_i^2) = 2^n \alpha^n \left( \prod_{i=1}^n x_i \right) \exp\left(-\alpha \sum_{i=1}^n x_i^2\right). \end{aligned}$$

Tar logaritmen:

$$l(x_1, x_2, \dots, x_n; \alpha) = \ln L(x_1, x_2, \dots, x_n; \alpha) = n \ln 2 + n \ln \alpha + \sum_{i=1}^n \ln x_i - \alpha \sum_{i=1}^n x_i^2.$$

Deriverer med hensyn på  $\alpha$  og setter lik 0:

$$\begin{aligned} \frac{\partial l(x_1, x_2, \dots, x_n; \alpha)}{\partial \alpha} &= 0 + \frac{n}{\alpha} + 0 - \sum_{i=1}^n x_i^2 = 0 \\ \frac{n}{\alpha} &= \sum_{i=1}^n x_i^2 \\ \alpha &= \frac{n}{\sum_{i=1}^n x_i^2}. \end{aligned}$$

Dvs. SME for  $\alpha$  er  $\alpha^* = \frac{n}{\sum_{i=1}^n x_i^2}$ , som er ulik  $\hat{\alpha}$ .  $\hat{\alpha}$  er dermed ikke SME for  $\alpha$ .

$$\text{Estimator } \hat{\mu} \text{ for } \mu: \hat{\mu} = \frac{\sqrt{\pi}}{2\sqrt{\hat{\alpha}}} = \frac{\sqrt{\pi} \sqrt{\sum_{i=1}^n X_i^2}}{2\sqrt{n-1}}.$$

Innsatt verdier blir  $\hat{\alpha} = 0.029$  og estimatet for  $\mu$  blir  $\hat{\mu} = \frac{\sqrt{\pi}}{2\sqrt{\hat{\alpha}}} = \frac{\sqrt{\pi}}{2\sqrt{0.029}} = 5.2$  uker.

c) Sannsynlighetsfordelingen for  $Y = X^2$  finner vi ved å bruke transformasjonsformelen.

La  $Y = u(X) = X^2$  slik at  $X = w(Y) = \sqrt{Y}$  (har at  $X > 0$ ). Dermed er

$$f_Y(y) = f_X(w(y)) |w'(y)| = 2\alpha \sqrt{y} \exp(-\alpha(\sqrt{y})^2) \left| \frac{1}{2\sqrt{y}} \right| = \alpha \exp(-\alpha y).$$

Dette er sannsynlighetstettheten i eksponensialfordelingen med forventning  $1/\alpha$ , og dermed har vi vist at  $Y$  er eksponensialfordelt.

Forventningsverdi for  $\hat{\alpha}$ :

$$E(\hat{\alpha}) = E\left(\frac{n-1}{\sum_{i=1}^n X_i^2}\right) = (n-1) \cdot E\left(\frac{1}{\sum_{i=1}^n X_i^2}\right) = (n-1) \cdot \frac{1}{(1/\alpha)(n-1)} = \frac{n-1}{n-1} \alpha = \alpha.$$

Her har vi brukt resultatet oppgitt i oppgaveteksten.

Siden  $E(\hat{\alpha}) = \alpha$ , er  $\hat{\alpha}$  forventningsrett.

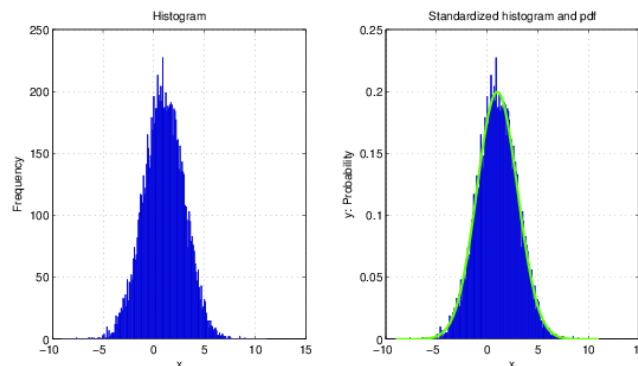
### Oppgave 3

Scriptet `run_confds.m` simulerer  $n$  data  $x_1, \dots, x_n$  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

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

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  $x_1, \dots, x_n$ , som f.eks. kan se slik ut

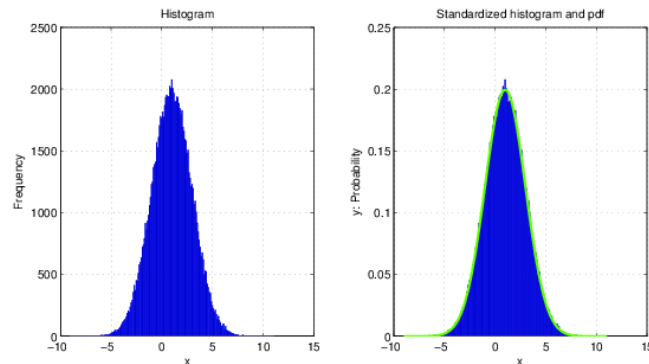


Figur 2: Histogram av  $n = 10000$  simulerte data fra  $N(1, 2^2)$

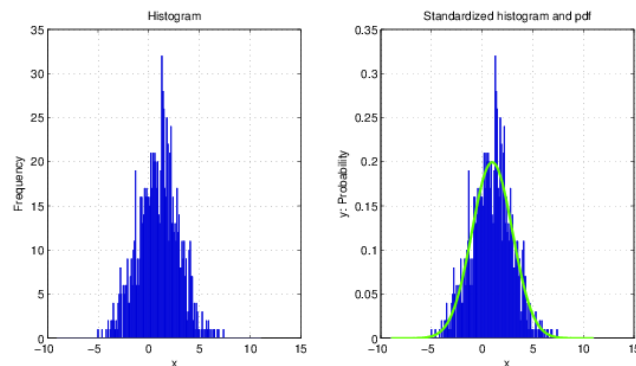
Histogrammet til høyre er standardisert, altså transformert slik at areal under histogram-søylene blir 1. I plottet er det i grønt også tegnet inn kurven for normalfordelingen med forventning 1 og standardavvik 2. Vi ser at de simulerte dataene overlapper normalfordelingen de kommer fra veldig bra. Dette 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 forventningsverdien som også ligger innenfor det estimerte konfidensintervallet  $[0.96591, 1.0434]$ .

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 forventningsverdi, som ligger innenfor konfidensintervallet, og overlappen mellom dataene og normalkurven er enda bedre.

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  $[0.83741, 1.0815]$ . 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).



Figur 3: Histogram av  $n = 100000$  simulerte data fra  $N(1, 2^2)$



Figur 4: Histogram av  $n = 1000$  simulerte data fra  $N(1, 2^2)$

Det estimerte konfidensintervallet 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 standardaviklet ikke varierer mye ser vi at faktoren  $\frac{\hat{\sigma}}{\sqrt{n}}$  går mot 0, altså blir konfidensintervallet 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 datamengden er så stor ( $n \geq 1000$ ) vil  $t$ -fordeling med  $n - 1$  frihetsgrader være tilnærmet lik standard normalfordeling.

#### Oppgave 4

- a) I dette punktet lar vi  $X$  være en bestemt pH-måling, og antar at  $X$  er normalfordelt med forventningsverdi  $\mu = 6.8$  og varians  $\sigma^2 = 0.060^2$ . Sannsynligheten for at resultatet

av målingen er under 6.74 er da

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

Videre er sannsynligheten for at resultatet av målingen ligger mellom 6.74 og 6.86 lik

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

Sannsynligheten for at avviket  $|X - \mu|$  overstiger 0.06 er

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

Den samme sannsynligheten kan også regnes ut som følger:

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

- b) Vi laster opp fila i Matlab og sjekker hvor stor andel av målingene som er lavere enn 6.74.

```
data=load('pH.txt')  
N= length(data) %Antall elementer i datasettet. Her: 53.
```

```
prob1=sum(data<6.74)/N
```

På samme måte undersøker vi hvor stor andel av målingene som er mellom 6.74 og 6.86, og andel målinger der forskjellen mellom målt verdi og  $\mu = 6.8$  er større enn 0.06.

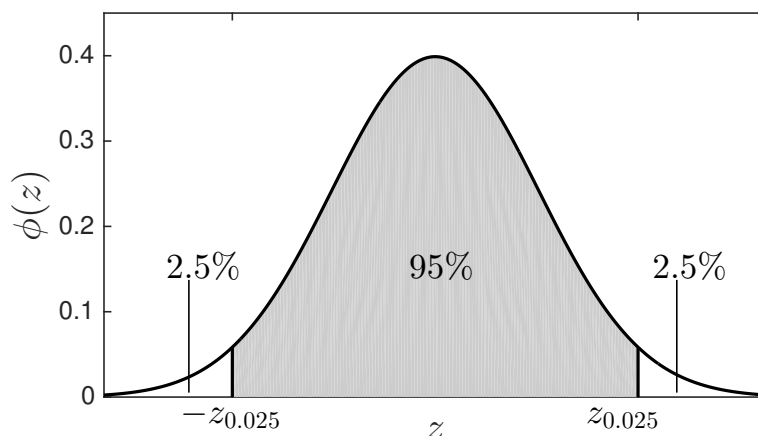
```
prob2=sum(data>6.74 & data < 6.86)/N
```

```
mu=6.8
```

```
prob3=sum(abs(data-mu)>0.06)/N
```

Datasettet gir svarene  $\text{prob1}=0.0755$ ,  $\text{prob2}=0.6415$  og  $\text{prob3}=0.3396$  som de empiriske sannsynlighetene for henholdsvis  $P(X < 6.74)$ ,  $P(6.74 < X < 6.86)$  og  $P(|X - \mu| < 0.06)$ . Datasettet gir sannsynligheter som er i nærheten av det vi fant i oppgave **a**). Hadde vi hatt enda flere målinger i datasettet, ville vi forventet at datasettet hadde gitt oss sannsynligheter som var enda nærmere de teoretiske verdiene fra **a**).

- c) Her er  $Y$  gjennomsnittet av de  $N$  uavhengige målingene  $X_1, X_2, \dots, X_N$ , som alle er normalfordelte med forventningsverdi  $\mu$  og varians  $\sigma^2$ . Siden  $Y$  er en lineærkombinasjon



Figur 5: Sannsynlighetstettheten  $\phi(z)$  til den standard normalfordelte stokastiske variabelen  $Z$ . Kvantilene  $\pm z_{0.025} = \pm 1.96$  er markert.

av normalfordelte tilfeldige variable, vet vi at  $Y$  selv er normalfordelt. Forventningsverdien og variansen til  $Y$  er henholdsvis

$$E(Y) = E\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{1}{N} \sum_{i=1}^N E(X_i) = \frac{1}{N} \cdot N\mu = \mu$$

og

$$\text{Var}(Y) = \text{Var}\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(X_i) = \frac{1}{N^2} \cdot N\sigma^2 = \frac{\sigma^2}{N},$$

slik at  $Y \sim N(\mu, \sigma^2/N)$ .

Den tilfeldige variabelen

$$Z = \frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}} = \frac{Y - \mu}{\sqrt{\sigma^2/N}}$$

er standard normalfordelt,  $Z \sim N(0, 1)$ . Et 95% konfidensintervall for  $\mu$  kan konstrueres ved å ta utgangspunkt i et tilsvarende konfidensintervall for  $Z$  (se figur 5), og så manipulere dette slik at  $\mu$  isoleres,

$$\begin{aligned} 0.95 &= 1 - 2 \cdot 0.025 \\ &= P(-z_{0.025} \leq Z \leq z_{0.025}) \\ &= P\left(-z_{0.025} \leq \frac{Y - \mu}{\sqrt{\sigma^2/N}} \leq z_{0.025}\right) \\ &= P\left(Y - z_{0.025} \sqrt{\frac{\sigma^2}{N}} \leq \mu \leq Y + z_{0.025} \sqrt{\frac{\sigma^2}{N}}\right). \end{aligned}$$

Vi bruker Matlab til å regne ut gjennomsnittet av de 53 målingene i `pH.txt` og regner ut et konfidensintervall for  $\mu$  basert på det spesifikke datasettet:

$N=53$

$\sigma=0.06$

$y=\text{mean}(\text{data})$

$\text{CI}=[y-1.96*\text{sqrt}(\sigma^2/N), y+1.96*\text{sqrt}(\sigma^2/N)]$

For dette datasettet er  $y = 6.8108$ . Setter vi inn den observerte verdien  $y = 6.8108$ , variansen  $\sigma^2 = 0.060^2$ , og 0.025-kvantilen til standard normalfordelingen  $z_{0.025} = 1.96$ , får vi konfidensintervallet:

$$\begin{aligned} \left[ y - z_{0.025} \sqrt{\frac{\sigma^2}{N}}, y + z_{0.025} \sqrt{\frac{\sigma^2}{N}} \right] &= \left[ 6.8108 - 1.96 \sqrt{\frac{0.060^2}{53}}, 6.8108 + 1.96 \sqrt{\frac{0.060^2}{53}} \right] \\ &= \underline{\underline{[6.7946, 6.8269]}}. \end{aligned}$$