Landmålingens fejlteori
Estimatering af μ og σ fra data
samt konfidens interval for μ
Lektion 5

Torben Tvedebrink - tvede@math.aau.dk
http://www.math.aau.dk/∼tvede/teaching/L4

Institut for Matematiske Fag
Aalborg Universitet

5. maj 2009

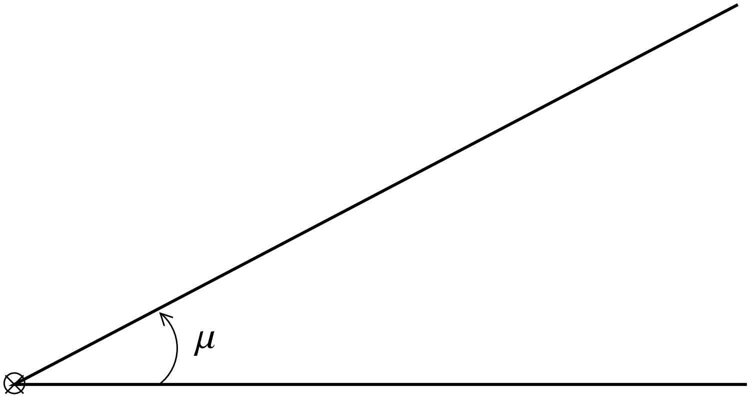
Statistisk model

I de fleste videnskaber antages en model for det fænomen som er under observation. Ofte bygger modellen på en række antagelser hvis påstande tidligere er vist valide.

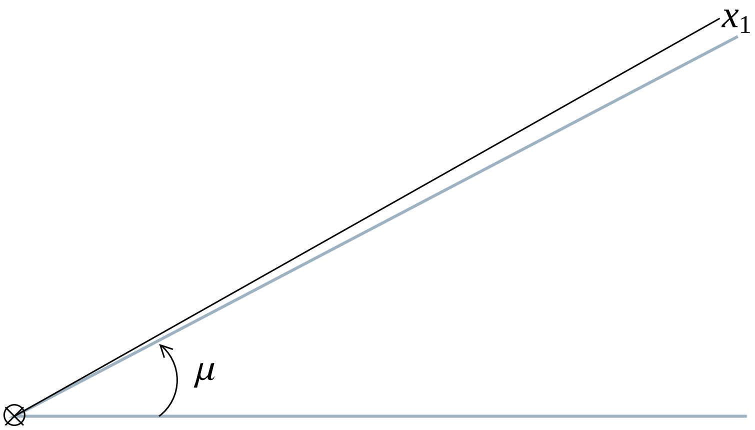
Model for vinkler i landmåling:
Vi antager at vinkler i landmåling er normalfordelte med den *sande* vinkel μ som middelværdi og spredning σ. Ydermere antages n gentagne måleforsøg $X_1, \ldots, X_n$ af samme vinkel at være uafhængige og identisk fordelte (iid),

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n.$$

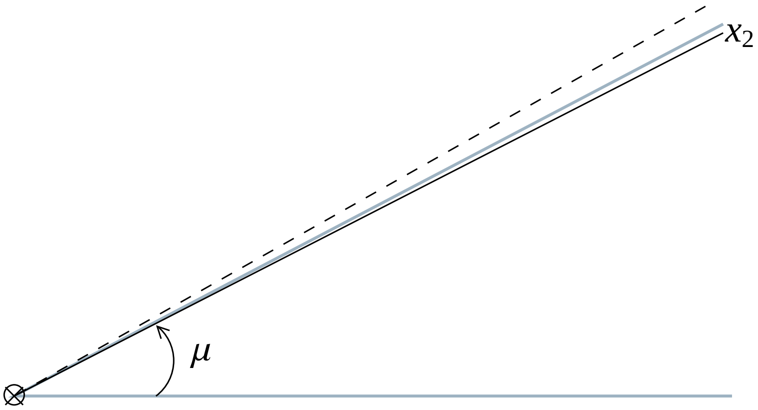
Model for vinkler



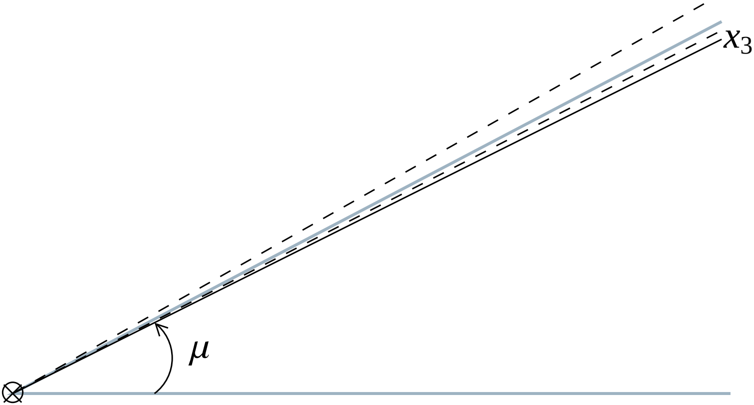
Model for vinkler



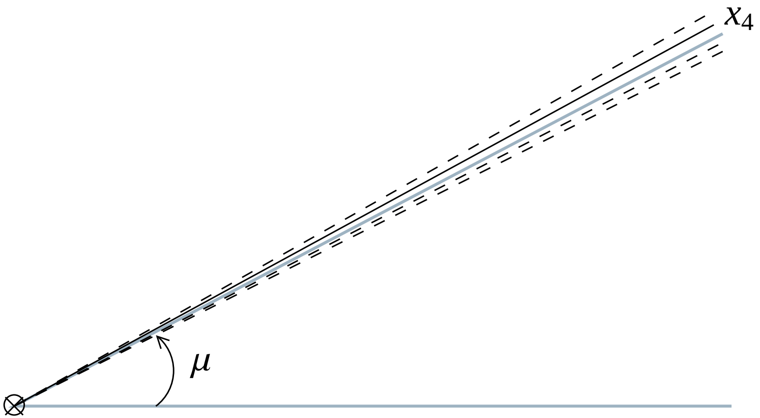
Model for vinkler



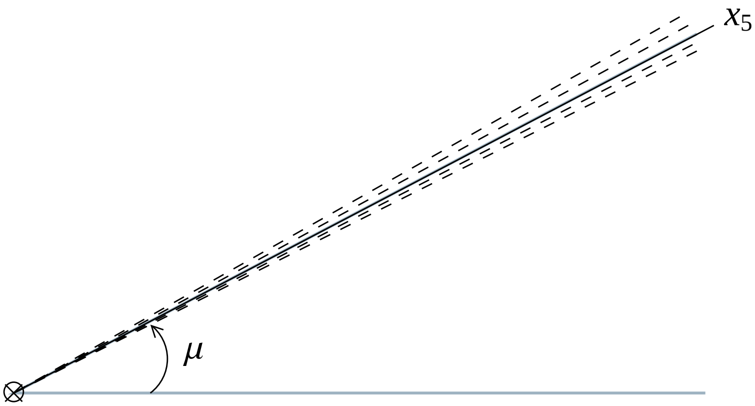
Model for vinkler



Model for vinkler



Model for vinkler



Observationer

Ved opmåling af en vinkel foretages n observationer $x_1, \ldots, x_n$ som er realisationer af de stokatiske variable $X_1, \ldots, X_n$.
Skematisk angives dette som,

$$\begin{array}{ccc} X_1 & \ldots & X_n \\ \downarrow & & \downarrow \\ x_1 & \ldots & x_n \end{array}$$

$(X_1, \ldots, X_n)$ kaldes en stikprøve fra normalfordelingen $\mathcal{N}(\mu, \sigma^2)$.

$(x_1, \ldots, x_n)$ kaldes en observeret stikprøve fra normalfordelingen $\mathcal{N}(\mu, \sigma^2)$.

Eksempel

Jf. Eksempel 1 fra noterne observeres følgende 10 satser af en vinkel.

Sats	x_i	Observation
1	x_1	164.508 gon
2	x_2	164.509 gon
3	x_3	164.511 gon
4	x_4	164.507 gon
5	x_5	164.510 gon
6	x_6	164.511 gon
7	x_7	164.517 gon
8	x_8	164.510 gon
9	x_9	164.514 gon
10	x_{10}	164.513 gon

Eksempel

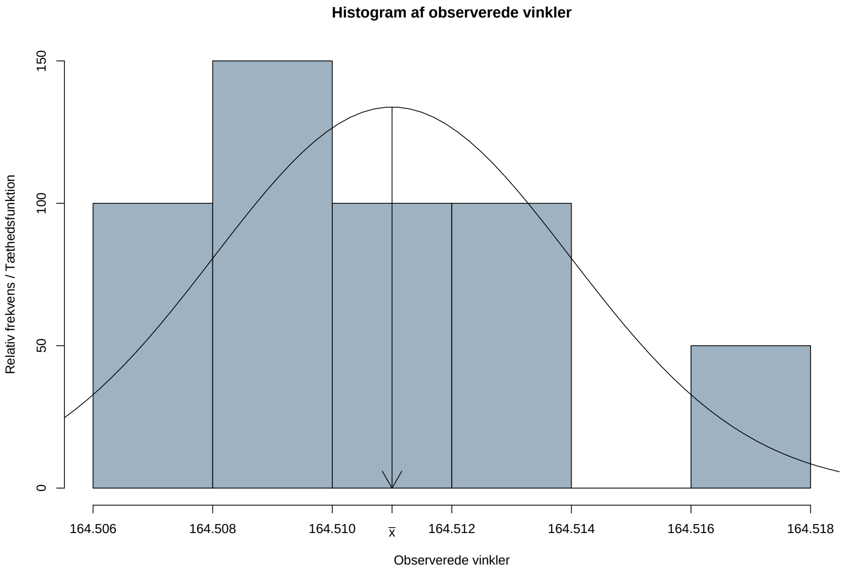
Jf. Eksempel 1 fra noterne observeres følgende 10 satser af en vinkel.

Sats	x_i	Observation
1	x_1	164.508 gon
2	x_2	164.509 gon
3	x_3	164.511 gon
4	x_4	164.507 gon
5	x_5	164.510 gon
6	x_6	164.511 gon
7	x_7	164.517 gon
8	x_8	164.510 gon
9	x_9	164.514 gon
10	x_{10}	164.513 gon

Dvs den observerede stikprøve, hvor $n = 10$, er

$$(x_1, x_2, \dots, x_{n-1}, x_n) = (164.508, 164.509, \dots, 164.514, 164.513).$$

Eksempel


Histogram af observerede vinkler
Relativ frekvens / Tæthedsfunktion
0
50
100
150
164.506
164.508
164.510
x̄
164.512
164.514
164.516
164.518
Observerede vinkler

Estimatorer

Som estimatorer for μ og σ^2 anvendes estimatorerne $\bar{X}$ og S^2. Disse er defineret som

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

Disse estimatorer kaldes centrale idet $\mathbb{E}(\bar{X}) = \mu$ og $\mathbb{E}(S^2) = \sigma^2$.

Estimater

Har vi observeret data kan vi estimere μ og σ^2 med $\bar{x}$ og s^2. Her udskiftes de stokatiske variable X_i i $\bar{X}$ og S^2 med de observerede x_i,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Både $\bar{X}$ og S^2 er stokastiske variable (transformationer af X_i'erne), mens $\bar{x}$ og s^2 er realisationer af disse,

$$\begin{array}{cccccc} X_1 & \ldots & X_n & & \bar{X} & S^2 \\ \downarrow & & \downarrow & & \downarrow & \downarrow \\ x_1 & \ldots & x_n & & \bar{x} & s^2 \end{array}$$

Eksempel - fortsat

Fra Eksempel 1 i noterne kan vi estimere μ med $\bar{x}$ og σ^2 med s^2.

$$\bar{x} = \frac{1}{10}(164.508 + 164.509 + \cdots + 164.514 + 164.513) = 164.511 \text{ gon}$$

$$\sum_{i=1}^{2} x_i^2 = 164.508^2 + 164.509^2 + \cdots + 164.514^2 + 164.513^2 = 270638.7 \text{ gon}^2$$

$$s^2 = \frac{1}{10-1}\left(270638.7 - 10 \cdot (164.511)^2\right) = (0.00298)^2 \text{ gon}^2$$

Størrelsen s^2 er et mål for nøjagtigheden af vores observationer.
Jo mindre desto mere nøjagtige er vores målinger.

Sætning 1

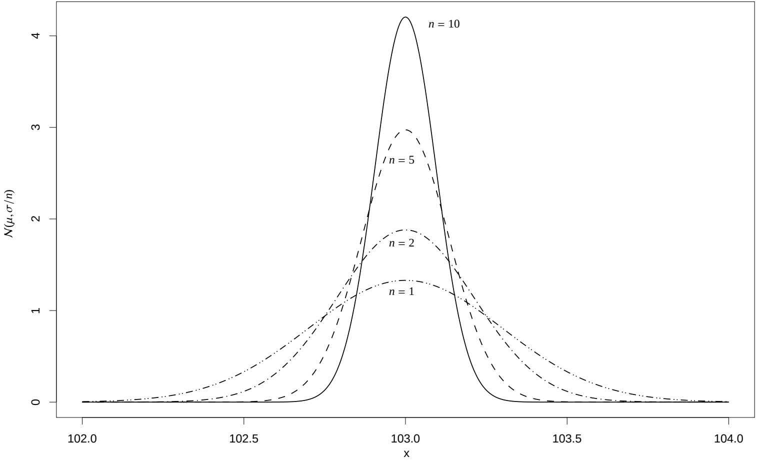
Antag $(X_1, \ldots, X_n)$ er en stikprøve fra en fordeling med middelværdi μ og varians σ^2. Da gælder

$$\mathbb{E}(\bar{X}) = \mu \quad \text{og} \quad \mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}.$$

Hvis $X_i \sim \mathcal{N}(\mu, \sigma^2)$ gælder der ligeledes $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Beviset følger fra sætninger gennemgået i Lektion 4.

Effekten af øget antal observationer



Estimatorer - fortsat*

Fra sidste sætning følger det at $\bar{X}$ er en central estimator for μ idet $\mathbb{E}(\bar{X}) = \mu$.

Nedenfor vises det at S^2 er en central estimator for σ^2, altså $\mathbb{E}(S^2) = \sigma^2$.

$$
\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)
\end{aligned}
$$

Estimatorer - fortsat*

Fra sidste sætning følger det at $\bar{X}$ er en central estimator for μ idet $\mathbb{E}(\bar{X}) = \mu$.

Nedenfor vises det at S^2 er en central estimator for σ^2, altså $\mathbb{E}(S^2) = \sigma^2$.

$$
\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)
\end{aligned}
$$

$$
\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 \quad \Leftrightarrow \quad \mathbb{E}(X_i^2) = \mathbb{V}(X_i) + \mathbb{E}(X_i)^2 = \sigma^2 + \mu^2
$$

Estimatorer - fortsat*

Fra sidste sætning følger det at $\bar{X}$ er en central estimator for μ idet $\mathbb{E}(\bar{X}) = \mu$.

Nedenfor vises det at S^2 er en central estimator for σ^2, altså $\mathbb{E}(S^2) = \sigma^2$.

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)
\end{aligned}$$

$$\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 \quad \Leftrightarrow \quad \mathbb{E}(X_i^2) = \mathbb{V}(X_i) + \mathbb{E}(X_i)^2 = \sigma^2 + \mu^2$$

$$\mathbb{V}(\bar{X}) = \mathbb{E}(\bar{X}^2) - \mathbb{E}(\bar{X})^2 \quad \Leftrightarrow \quad \mathbb{E}(\bar{X}^2) = \mathbb{V}(\bar{X}) + \mathbb{E}(\bar{X})^2 = \sigma^2/n + \mu^2$$

Estimatorer - fortsat*

Dvs. vi har indtil videre:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)$$

$$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2 \qquad \mathbb{E}(\bar{X}^2) = \sigma^2/n + \mu^2$$

Indsættes udtryk for middelværdi af X_i og $\bar{X}$ i $\mathbb{E}(S^2)$ får vi:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right)$$

Estimatorer - fortsat*

Dvs. vi har indtil videre:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)$$

$$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2 \qquad \mathbb{E}(\bar{X}^2) = \sigma^2/n + \mu^2$$

Indsættes udtryk for middelværdi af X_i og $\bar{X}$ i $\mathbb{E}(S^2)$ får vi:

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \frac{n}{n-1}\sigma^2 + \frac{n}{n-1}\mu^2 - \frac{n}{n-1}\frac{\sigma^2}{n} - \frac{n}{n-1}\mu^2\end{aligned}$$

Estimatorer - fortsat*

Dvs. vi har indtil videre:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)$$

$$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2 \qquad \mathbb{E}(\bar{X}^2) = \sigma^2/n + \mu^2$$

Indsættes udtryk for middelværdi af X_i og $\bar{X}$ i $\mathbb{E}(S^2)$ får vi:

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n}+\mu^2\right)\\ &= \frac{n}{n-1}\sigma^2 + \frac{n}{n-1}\mu^2 - \frac{n}{n-1}\frac{\sigma^2}{n} - \frac{n}{n-1}\mu^2\\ &= \frac{n\sigma^2-\sigma^2}{n-1}\end{aligned}$$

Estimatorer - fortsat*

Dvs. vi har indtil videre:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\bar{X}^2)$$

$$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2 \qquad \mathbb{E}(\bar{X}^2) = \sigma^2/n + \mu^2$$

Indsættes udtryk for middelværdi af X_i og $\bar{X}$ i $\mathbb{E}(S^2)$ får vi:

$$\begin{aligned}
\mathbb{E}(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= \frac{n}{n-1}\sigma^2 + \frac{n}{n-1}\mu^2 - \frac{n}{n-1}\frac{\sigma^2}{n} - \frac{n}{n-1}\mu^2 \\
&= \frac{n\sigma^2 - \sigma^2}{n-1} \\
&= \sigma^2.
\end{aligned}$$

Estimatorer - fortsat

Havde vi istedet divideret med n i udtrykket for S^2 ville middelværdien være

$$\mathbb{E}(S^2) = \frac{n\sigma^2 - \sigma^2}{n} = \frac{(n-1)}{n}\sigma^2 \neq \sigma^2.$$

En intuitiv forklaring på at vi dividerer med $n-1$ er at vi "forbruger" én observation på at estimere μ med $\bar{x}$.

Fx. har vi observeret x_1 og x_2 er $\bar{x} = (x_1 + x_2)/2$. Dvs. ud fra x_1 og $\bar{x}$ kan vi bestemme $x_2 = 2\bar{x} - x_1$ og der er altså indført "bindinger" på den n'te observation.

Estimatorer - kendt middelværdi μ

I situationer hvor vi kender μ (fx. på en øvelsesbane hvor sande længder og vinkler er kendt) bruger vi derfor $\hat{s}$ til at estimere målingernes nøjagtighed,

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Estimatorer - kendt middelværdi μ

I situationer hvor vi kender μ (fx. på en øvelsesbane hvor sande længder og vinkler er kendt) bruger vi derfor $\hat{s}$ til at estimere målingernes nøjagtighed,

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

I disse situationer er $\hat{S}^2$ et centralt estimat for σ^2. Dvs:

$$\mathbb{E}(\hat{S}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i^2 + \mu^2 - 2\mu X_i)\right) = \sigma^2.$$

Kan bevises på tilsvarende måde som $\mathbb{E}(S^2) = \sigma^2$.

Konfidensinterval for μ

Vores modelantagelse giver at $(X_1, \dots, X_n)$ er uafhængige og identiske fordelte, $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \dots, n$. Ydermere antages det her at variansen σ^2 er kendt.

Dette medfører at $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Jf. slide 18 (3. Lektion) gælder der for $Y \sim \mathcal{N}(\mu, \sigma^2)$ at

$$P(\mu - 1.96\sigma \leq Y \leq \mu + 1.96\sigma) = 0.95$$

Tilsvarende resultat gælder for $\bar{X}$ hvor spredningen blot er $\sigma/\sqrt{n}$,

$$P(\mu - 1.96\tfrac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96\tfrac{\sigma}{\sqrt{n}}) = 0.95$$

Konfidensinterval - fortsat

For uligheden inde i $P()$ har vi,

$$\begin{array}{rcccl} \mu - 1.96\frac{\sigma}{\sqrt{n}} & \leq & \bar{X} & \leq & \mu + 1.96\frac{\sigma}{\sqrt{n}} \\ -\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} & \leq & -\mu & \leq & -\bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} & \geq & \mu & \geq & \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \end{array}$$

Uligheden skrives nu som sædvanligt

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Konfidensinterval - fortsat

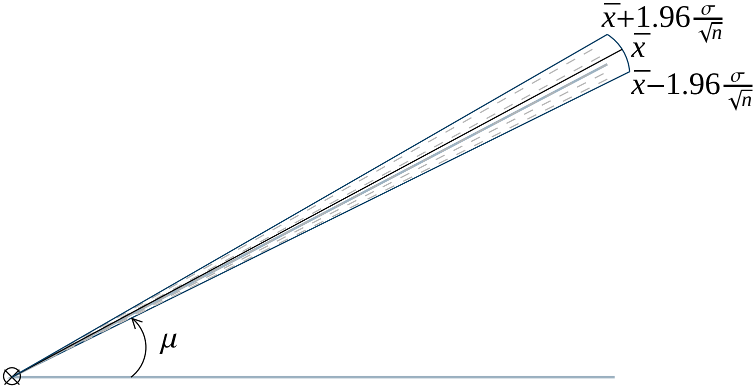
Dette er ensbetydende med at

$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Hvor det stokastiske ligger i $\bar{X}$. Dvs det er intervallet som er stokatisk.

"Sandsynligheden for at $\bar{X}$ antager en værdi $\bar{x}$ så μ ligger i intervallet $[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\,;\, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ er 0.95".

Konfidensinterval - grafisk



Konfidensinterval - fortolkning

Antag vi observerer de n stokatiske variable k gange, dvs. vi får k observationsrækker med n tal.

$$\begin{aligned}
1 &: x_{1,1}, x_{1,2}, \ldots, x_{1,n} &\longrightarrow\quad & \bar{x}_1 \\
2 &: x_{2,1}, x_{2,2}, \ldots, x_{2,n} &\longrightarrow\quad & \bar{x}_2 \\
&\vdots \\
k &: x_{k,1}, x_{k,2}, \ldots, x_{k,n} &\longrightarrow\quad & \bar{x}_k
\end{aligned}$$

Hermed fås k middelværdi estimater $\bar{x}_1, \ldots, \bar{x}_k$ og k tilhørende konfidensintervaller. For k stor kan vi forvente at 95% af intervallerne indeholder μ.

Eksempel - fortsat

Antag at vi kendte variansen i vores eksempel med 10 observerede vinkelmålinger. Det oplyses at $\sigma^2 = 0.002^2$. Vi kan da bestemme et 95% konfidensinterval for μ, hvor $\bar{x} = 164.511$ fra tidligere:

$$\left[164.511 - 1.96\frac{0.002}{\sqrt{10}}\ ;\ 164.511 + 1.96\frac{0.002}{\sqrt{10}}\right] = [164.5098\ ;\ 164.5122]$$

Per konstruktion ligger $\bar{x}$ altid midt i intervallet. Længden på intervallet er et udtryk for nøjagtigheden, hvor et kortere interval indikerer at μ er bedre estimeret end et længere.

Eksempel

Der foretages 20 gange 10 opmålinger af en længde på 118.12 m. Det antages at der er en varians på observationerne på 0.02 m^2. Figuren viser de 20 konfidensintervaller for hver forsøgsrække.