Landmålingens fejlteori
Lektion 2
Sandsynlighedsintervaller
Estimation af μ
Konfidensinterval for μ

Rasmus Waagepetersen - rw@math.aau.dk

Institut for Matematiske Fag
Aalborg Universitet

En stokastisk variabel er en variabel, som vi tillægger sandsynligheder - vha. tætheds- eller fordelingsfunktion.

Ex. ligelig fordelt stokastisk variabel på $[a,b]$:

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{ellers} \end{cases} \quad F(x) = \int_{-\infty}^{x} f(z)\mathrm{d}z = \begin{cases} 0 & x < a \\ \frac{x}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$

Middelværdi

$$\mu = EX = \int_a^b x \frac{1}{b-a}\mathrm{d}x = (b+a)/2$$

Varians

$$\sigma^2 = \mathbb{V}\mathrm{ar}X = \int_a^b (x-\mu)^2 \frac{1}{b-a}\mathrm{d}x = EX^2 - \mu^2 = \frac{(b-a)^2}{12}$$

Spredning

$$\sigma = \sqrt{\mathbb{V}\mathrm{ar}X} = \frac{b-a}{\sqrt{12}}$$

Definition: Normalfordelingen

En stokastisk variabel X med **tæthedsfunktion**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right), \quad x \in \mathbb{R},$$

siges at være **normalfordelt** med middelværdi μ og varians σ^2, hvor μ og σ er reelle tal og $\sigma > 0$.

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$.

Fordelingsfunktion

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(z-\mu)^2}{2\sigma^2} \right) \mathrm{d}z$$

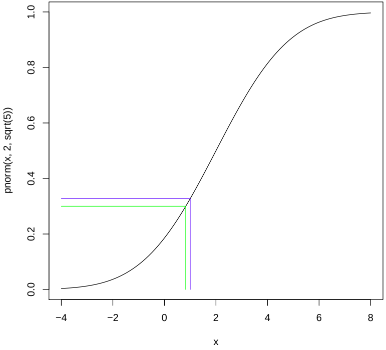
har ikke 'lukket' udtryk.

Sandsynligheder og fraktiler udregnes mest bekvemt vha. computerprogram - f.eks. `matlab`.

Eksempel $X \sim N(2{,}5)$.

Beregn $P(X \leq 1)$ samt $x_{0.3} = F^{-1}(0{,}3)$:

```
>> normcdf(1,2,sqrt(5))
ans =
    0.3274
>> norminv(0.3,2,sqrt(5))
ans =
   0.8274
```



Repetition: Standard-normalfordelingen

Definition: Standard-normalfordelingen

Fordelingen $\mathcal{N}(0,1)$ kaldes **standard normalfordelingen**. Typisk noteres standard-normal fordelte variable Z.

Historisk set nyttig, da nok at tabellægge standardnormalfordelingen for at udregne sandsynligheder for alle andre normalfordelinger:

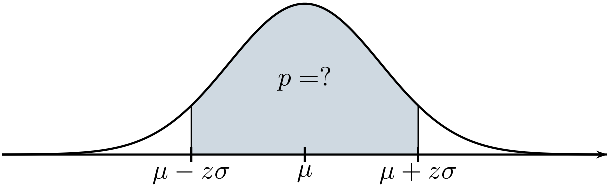
$$X \sim \mathcal{N}(\mu,\sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Sandsynlighedsintervaller (Sætning 8)

Lad $X \sim \mathcal{N}(\mu, \sigma^2)$, dvs. $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Find sandsynligheden for at X ligger højst z standardafvigelser fra middelværdien μ. Dvs. find:

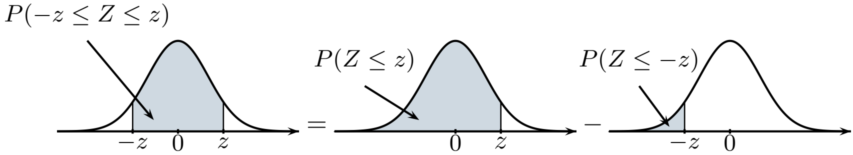
$$p \ = P(\mu - z\sigma < X < \mu + z\sigma)$$



Sandsynlighedsintervaller

Sandsynligheden for at X ligger højst z standardafvigelser fra middelværdien μ er:

$$
\begin{aligned}
p = P(\mu - z\sigma < X < \mu + z\sigma) &= P(-z < \frac{X-\mu}{\sigma} < z) \\
&= P(-z < Z < z) \\
&= P(Z < z) - P(Z < -z) \\
&= \Phi(z) - \Phi(-z) \\
&= \Phi(z) - (1 - \Phi(z)) \\
&= 2\Phi(z) - 1.
\end{aligned}
$$



Bemærk: middelværdien μ og spredningen σ indgår *ikke*! Spredning nyttig enhed for måling af usikkerhed.

Sandsynlighedsintervaller: Eksempel

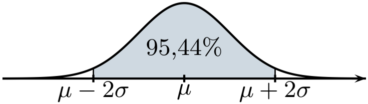
Antag $X \sim \mathcal{N}(\mu,\sigma^2)$. Hvad er sandsynligheden for at X ligger højst 2 standardafvigelser fra middelværdien?

$$p = P(\mu - 2\sigma < X < \mu + 2\sigma) = 2\Phi(2) - 1$$

Vha. `matlab`/ normalfordelingstabellen/ finder vi $\Phi(2) = 0{,}9772$. Dvs.

$$p = 2 \cdot 0{,}9772 - 1 = 0{,}9544.$$

Dvs. der er 95,44% sandsynlighed for at en normalfordelt SV ligger indenfor 2 standardafvigelser fra middelværdien.



Sandsynlighedsintervaller: Eksempler

Antag $X \sim \mathcal{N}(\mu,\sigma^2)$. Givet en sandsynlighed p find z så

$$P(\mu - z\sigma \leq X \leq \mu + z\sigma) = p.$$

Vi har set, at dette svarer til at løse $p = 2\Phi(z) - 1$. Isolerer vi $\Phi(z)$ får vi

$$\Phi(z) = \frac{p+1}{2} \quad \Leftrightarrow \quad z = \Phi^{-1}\left(\frac{p+1}{2}\right).$$

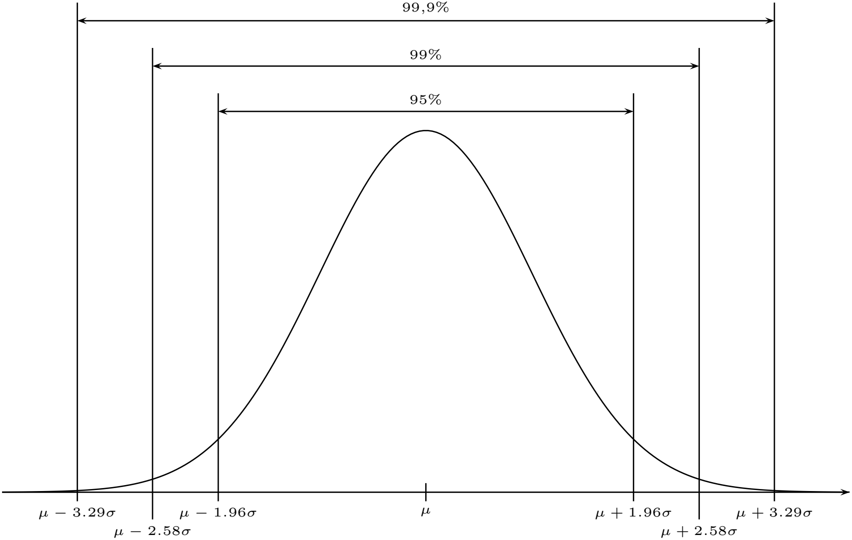
Antag vi vil finde z, så intervallet indeholder X med 99% sandsynlighed, dvs. $p = 0{,}99$. Da er $z = \Phi^{-1}(\frac{0{,}99+1}{2}) = \Phi^{-1}(0{,}995)$. Fra Matlab får vi

```
>> norminv(0.995)
ans =
    2.5758
```

Dvs.

$$P(\mu - 2{,}58\sigma \leq X \leq \mu + 2{,}58\sigma) = 0{,}99.$$

Tæthedsfunktion



Uafhængige stokastiske variable

To stokastiske variable X_1 og X_2 uafhængige hvis viden om X_1's værdi ikke indvirker på sandsynlighedsfordelingen af X_2 (og omvendt):

$$P(X_2 \leq x_2 \,|\, X_1 \in [a,b]) = P(X_2 \leq x_2)$$

Venstresiden læses som sandsynligheden for $X_2 \leq x_2$ *givet* at $X_1 \in [a,b]$.

Gentagne målinger skal gerne være uafhængige: hvis eks. måling X_1 er for stor ville det være uheldigt, hvis det medførte en forøget sandsynlighed for, at X_2 også bliver for stor.

(Definition 4 giver en stringent, men ikke helt intuitiv definition på uafhængighed)

Linearkombination

Sætning: Middelværdi af linearkombination

Hvis $X_1, X_2, \ldots, X_n$ er stokastiske variable med middelværdier

$$\mathbb{E}(X_1) = \mu_1, \mathbb{E}(X_2) = \mu_2, \ldots, \mathbb{E}(X_n) = \mu_n,$$

og $a_0, a_1, \ldots, a_n \in \mathbb{R}$, så gælder

$$\begin{aligned}\mathbb{E}\big(a_0 + a_1 X_1 + \cdots + a_n X_n\big) &= a_0 + a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n) \\ &= a_0 + a_1 \mu_1 + \cdots + a_n \mu_n\end{aligned}$$

Bemærk: Sætningen kræver *ikke* at $X_1, X_2, \ldots, X_n$ er uafhængige!

Anvendelse: omskrivning af udtryk for varians

Definitionen på varians kan omskrives til

$$\sigma^2 = \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2) - \mu^2,$$

Linearkombination

Sætning: Varians af linearkombination af uafhængige SV

Hvis $X_1, X_2, \ldots, X_n$ er **uafhængige** stokastiske variable med varianser

$$\mathbb{V}\mathrm{ar}(X_1) = \sigma_1^2, \ \mathbb{V}\mathrm{ar}(X_2) = \sigma_2^2, \ldots, \mathbb{V}\mathrm{ar}(X_n) = \sigma_n^2,$$

og a_0,$a_2, \ldots, a_n$ er reelle konstanter, så gælder

$$\mathbb{V}\mathrm{ar}\big(a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n\big) = a_1^2\mathbb{V}\mathrm{ar}(X_1) + a_2^2\mathbb{V}\mathrm{ar}(X_2) + \cdots + a_n^2\mathbb{V}\mathrm{ar}(X_n) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.$$

NB: vi skal senere se (lektion 5) at det faktisk er nok, at $X_1, \ldots, X_n$ er *ukorrelerede*.

Sum af normalfordelte SV

Sætning: Sum af normalfordelte SV

Hvis $X_1, X_2, \dots X_n$ er normalfordelte stokastiske variable,

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{for} \quad i = 1,2,\dots,n,$$

og $a_0,\dots,a_n$ er konstanter, så er summen

$$a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

normalfordelt.

Fra de foregående resultater fås at summen har middelværdi

$$a_0 + a_1 \mu_1 + a_2 \mu_2 + \cdots + a_n \mu_n$$

og hvis $X_1, X_2, \dots X_n$ er uafhængige (eller blot ukorrelerede) er variansen af summen

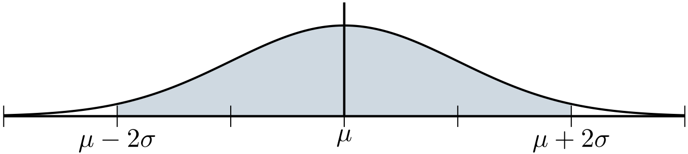
$$a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_n^2 \sigma_n^2$$

Statistisk model

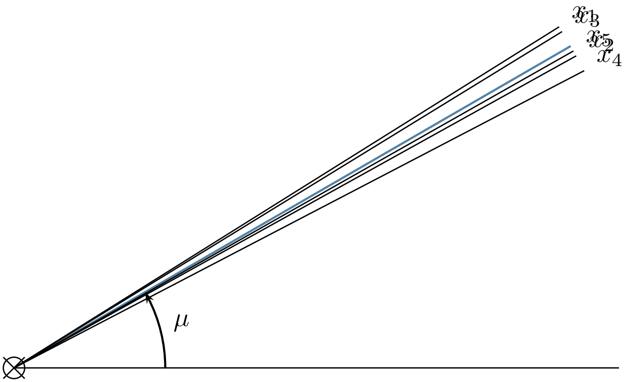
Model for vinkler i landmåling:

Vi antager, at vinkler i landmåling er normalfordelte med den *sande* vinkel μ som middelværdi og spredning σ. Ydermere antages n gentagne målinger $X_1, \ldots, X_n$ 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



Observationer

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

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

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

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

Eksempel

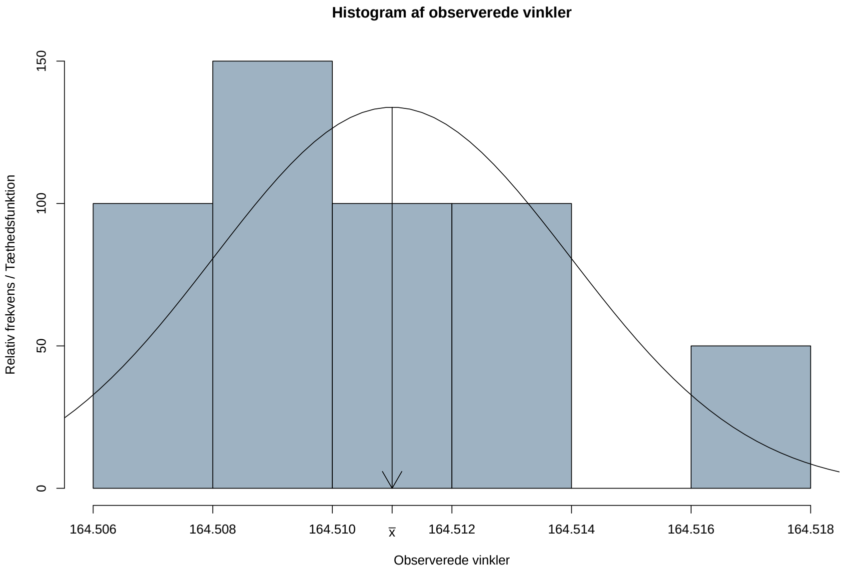
Jf. eksempel fra noterne observeres følgende 10 satser af en vinkel.

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

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

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

Estimator og estimat

Som estimator for μ anvendes gennemsnittet $\bar{X}$, der er defineret som

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

Har vi observeret data kan vi estimere μ med $\bar{x}$. Her udskiftes de stokastiske variable X_i i $\bar{X}$ ud med de observerede x_i,

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

Bemærk: $\bar{X}$ er en stokastisk variabel (en transformation af X_i'erne), mens $\bar{x}$ er en realisation af $\bar{X}$,

$$\begin{array}{ccccc} X_1 & \ldots & X_n & & \bar{X} \\ \downarrow & & \downarrow & & \downarrow \\ x_1 & \ldots & x_n & & \bar{x} \end{array}$$

Eksempel - fortsat

For eksemplet kan vi estimere μ med $\bar{x}$:

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

Egenskaber ved $\bar{X}$

Sætning: Middelværdi og varians for $\bar{X}$

Antag $X_1, \dots, X_n$ er uafhængige stokastiske variable med fælles middelværdi μ og varians σ^2. Da gælder

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

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

Estimatoren $\bar{X}$ kaldes en *central estimator* for μ, idet $\mathbb{E}(\bar{X}) = \mu$.

Estimatet $\bar{x}$ kaldes et *centralt estimat* for μ.

Bemærk: stort n giver lille varians/stor præcision !

Egenskaber for $\bar{X}$: Bevis

Vi har antaget at $X_1,\ldots,X_n$ er indbyrdes uafhængige og har *samme* middelværdi μ og samme varians σ^2:

$$\mathbb{E}(X_i) = \mu \quad , \quad \mathbb{V}\mathrm{ar}(X_i) = \sigma^2 \quad i = 1,\ldots,n.$$

Bemærk $\frac{1}{n}X_i$ har middelværdi $\frac{1}{n}\mu$ og varians $\frac{1}{n^2}\sigma^2$. Dermed har

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

middelværdi

$$\frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = n\frac{1}{n}\mu = \mu$$

og varians

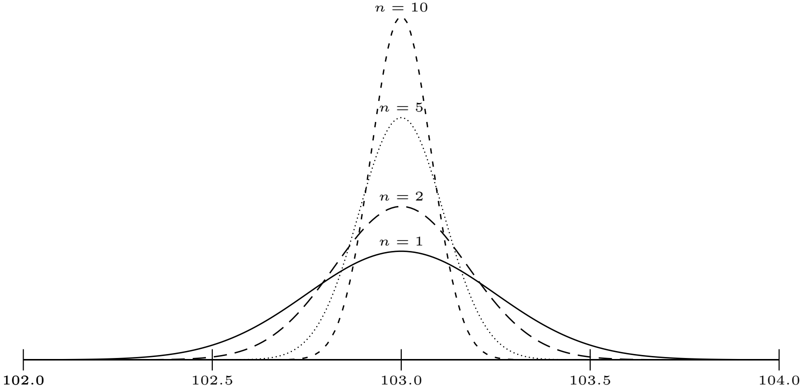
$$\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = n\frac{1}{n^2}\sigma^2 = \frac{1}{n}\sigma^2$$

Hvis yderligere de enkelte X_i er normalfordelte gælder $\frac{1}{n}X_i \sim N(\frac{1}{n}\mu, \frac{1}{n^2}\sigma^2)$ og dermed er også $\bar{X}$ normalfordelt.

(husk, sum af normalfordelte er selv normalfordelt).

Effekten af øget antal observationer

Fordelingen af gennemsnittet $\bar{X}$ for forskellige antal observationer (n).



Bemærk: Jo større n jo større sandsynlighed for at $\bar{X}$ ligger "tæt på" μ.

Konfidensinterval for μ

Gennemsnittet $\bar{x}$ er et estimat for μ. Hvor præcist er dette estimat?

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

Fra teorien om sandsynlighedsintervaller har vi for $X \sim \mathcal{N}(\mu,\sigma^2)$:

$$P(\mu - 1{,}96\sigma \le X \le \mu + 1{,}96\sigma) = 0{,}95 \ .$$

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

Heraf følger, at for $\bar{X}$ gælder:

$$P(\mu - 1{,}96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95 \ .$$

Konfidensinterval - fortsat

Sandsynligheden for forrige slide kan nu omskrives, så μ “isoleres”:

$$
\begin{array}{lccccc}
P(& \mu - 1.96\frac{\sigma}{\sqrt{n}} & \leq & \bar{X} & \leq & \mu + 1.96\frac{\sigma}{\sqrt{n}} &) = 0.95 \\
P(& -\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} & \leq & -\mu & \leq & -\bar{X} + 1.96\frac{\sigma}{\sqrt{n}} &) = 0.95 \\
P(& \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} & \geq & \mu & \geq & \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} &) = 0.95 \\
P(& \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} & \leq & \mu & \leq & \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} &) = 0.95
\end{array}
$$

Den sidste sandsynlighed har følgende fortolkning:

"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".

eller

"Der er 95% sandsynlighed for at det stokastiske interval $[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\,;\, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$ indeholder μ."

Konfidensinterval - fortsat

Vi kan nu definere et konfidensinterval

Definition: Konfidensinterval

Antag $X_1, \ldots, X_n$ er en stikprøve af uafhænige observationer fra $\mathcal{N}(\mu,\sigma^2)$, og $\bar{X}$ er gennemsnittet af denne stikprøve. Da er et 95% konfidensinterval for μ givet ved

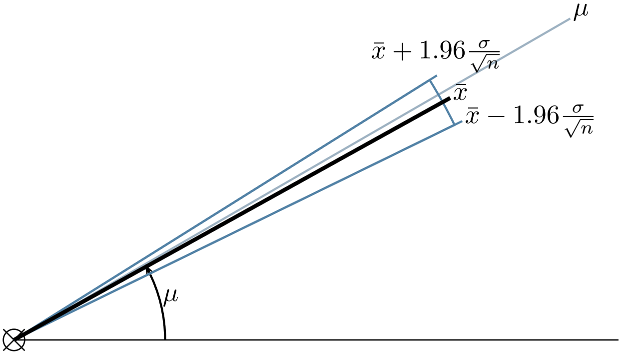
$$\bar{X} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}.$$

Fortolkning: Vi er 95% sikre på, at intervallet $\bar{X} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}$ indeholder den sande middelværdi μ.

Omvendt: for en given konkret realisation $\bar{x} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}$ vil der være sandsynlighed enten 0 eller 1 for at μ er i intervallet.

Konfidensintervallet har forskellig fortolkning før og og efter data er observeret !

Konfidensinterval - grafisk



Konfidensinterval - fortolkning

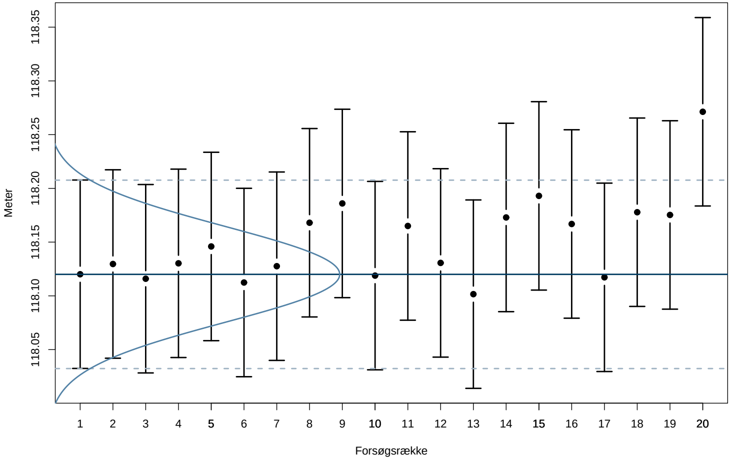
Antag vi observerer de n stokastiske variable k gange, dvs. vi får k observationsrækker med n tal.

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

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

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.



Praktisk brug af konfidensinterval

In praktisk brug vil vi ofte agere, som at den sande middelværdi er en af værdierne i 95% konfidensintervallet.

Det kan meget vel være forkert for et givet observeret konfidensinterval.

Men i det lange løb tager vi kun fejl i 5% af tilfældene.

Vil vi have større sikkerhed kan vi benytte f.eks. 99% konfidensintervaller (skift faktoren 1.96 ud med 2.58).

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 (kort interval=høj præcision)

Ikke normalfordelte data - er alt tabt ?

'Magisk' resultat (centrale grænseværdisætning, CLT)

Centrale grænseværdisætning

Hvis $X_1, \ldots, X_n$ er uafhængige stokastiske variable med samme middelværdi μ og varians σ^2 så gælder

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

når n 'stor'

Vores konstruktion af konfidensinterval benyttede blot, at $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

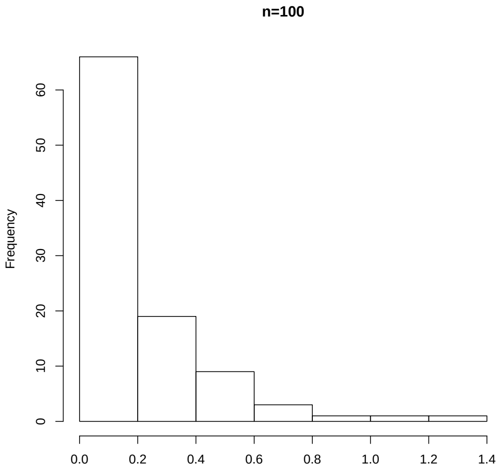
Dvs. selv for ikke-normalfordelte målinger vil konfidensintervallet stadig give god mening.

Illustration af CLT

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

hvor $x_{i,j}$ realisationer af Gamma-fordelte stokastiske variable.

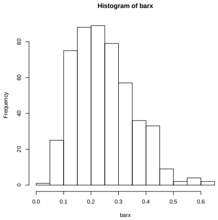
Histogram af første stikprøve $n = 100$



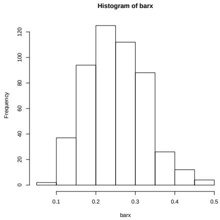
Observationer langt fra normalfordelte !

Histogram af 500 gennemsnit $\bar{x}_1, \ldots \bar{x}_{500}$

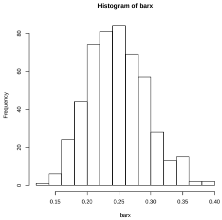
$n = 5$



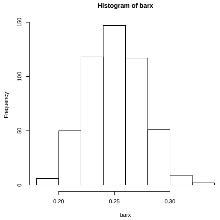
$n = 10$



$n = 30$



$n = 100$



Gennemsnit af mange observationer er normalfordelt !