Landmålingens fejlteori
Lektion 3

Rasmus Waagepetersen - rw@math.aau.dk

Institut for Matematiske Fag
Aalborg Universitet

Repetition: Middelværdi og Varians

Sætning: Middelværdi og varians for linearkombinationer

Lad $X_1, X_2, \ldots, X_n$ være stokastiske variable. Da gælder

$$\mathbb{E}\big(a_0 + a_1X_1 + \cdots + a_nX_n\big) = a_0 + a_1\mathbb{E}(X_1) + \cdots + a_n\mathbb{E}(X_n)$$

Hvis $X_1, X_2, \ldots, X_n$ desuden er uafhængige gælder

$$\mathbb{V}\mathrm{ar}\big(a_0 + a_1X_1 + \cdots + a_nX_n\big) = a_1^2\mathbb{V}\mathrm{ar}(X_1) + \cdots + a_n^2\mathbb{V}\mathrm{ar}(X_n)$$

Hvis X_i'erne er normalfordelte, så er summen $a_0 + a_1X_1 + \cdots + a_nX_n$ også normalfordelt.

Estimation af middelværdi

Antag $X_1, \ldots, X_n$ er uafhængige stokastiske variable med middelværdi μ og varians σ^2.

Vi estimerer μ ved gennemsnittet

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

Middelværdi og varians af $\bar{X}$:

$$\mathbb{E}\bar{X} = \mu \quad \mathbb{V}\mathrm{ar}\bar{X} = \frac{\sigma^2}{n}$$

Gentagne målinger og (u)afhængighed

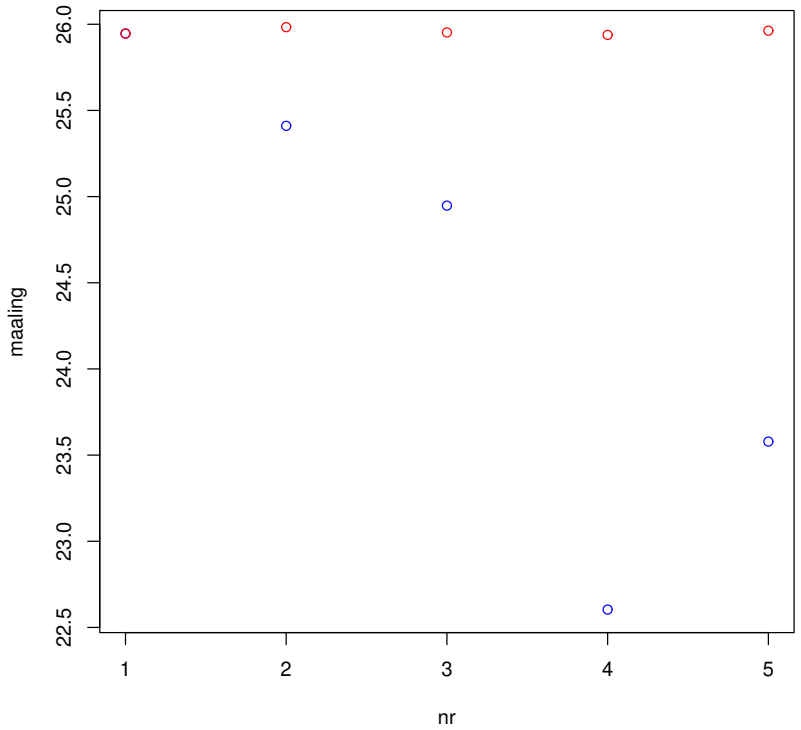
En gruppe med 5 medlemmer skal estimere en afstand

To strategier for at få bedre præcision

- Søren måler fem gange. Gennemsnit $\bar{x} = 26.0$
- Hver af gruppemedlemmerne måler en gang. Gennemsnit $\bar{x} = 24.5$

Målinger (Søren rød, gruppe blå):

Hvilket af estimaterne skal vi bruge ?



Afhængige vs. uafhængige målinger

Lad X og Y repræsentere målinger foretaget af Bent og Børge uafhængigt af hinanden og begge med varians σ^2 (og samme middelværdi μ).

Antag Bent måler en gang til, hvor hans anden måling Z er påvirket af hans resultat for den første måling, dvs.

$$Z = X + \nu$$

hvor ν er uafhængig af X med en (lille) varians ω^2 og middelværdi nul.

Dermed er variansen på $(X+Y)/2$ lig $\sigma^2/2$ mens variansen på $(X+Z)/2$ er $\sigma^2 + \omega^2/4$. Dvs. langt den bedste præcision med uafhængige målinger.

Omvendt: $\mathbb{V}\mathrm{ar}(X-Y) = 2\sigma^2$ mens $\mathbb{V}\mathrm{ar}(X-Z) = \omega^2$, hvad der (fejlagtigt) kunne forlede landinspektørerne til at tro, at Børges målinger var mest præcise.

Bias og varians

Vi modellerer en måling som

$$X = \mu + \epsilon$$

hvor fejlen har middelværdi nu, $\mathbb{E}\epsilon = 0$ og varians σ^2.

Hvad hvis der er en systematisk fejl (bias) så $\mathbb{E}\epsilon = b \neq 0$?

For gennemsnit $\bar{X}$ af n uafhængige målinger fordelt som X gælder:

$$\mathbb{E}\bar{X} = \mu + b \quad \mathbb{V}\mathrm{ar}\bar{X} = \frac{\sigma^2}{n}$$

Dvs. variansen går mod nul når n øges, men fejlen i bestemmelsen af μ forsvinder ikke !

Vi bliver meget sikre på et forkert resultat !

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

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

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

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

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

Konfidensinterval - fortsat

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

$$
\begin{aligned}
P(\quad \mu - 1.96\tfrac{\sigma}{\sqrt{n}} &\leq \bar{X} \leq \mu + 1.96\tfrac{\sigma}{\sqrt{n}} \quad) = 0.95 \\
P(\quad -\bar{X} - 1.96\tfrac{\sigma}{\sqrt{n}} &\leq -\mu \leq -\bar{X} + 1.96\tfrac{\sigma}{\sqrt{n}} \quad) = 0.95 \\
P(\quad \bar{X} + 1.96\tfrac{\sigma}{\sqrt{n}} &\geq \mu \geq \bar{X} - 1.96\tfrac{\sigma}{\sqrt{n}} \quad) = 0.95 \\
P(\quad \bar{X} - 1.96\tfrac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{X} + 1.96\tfrac{\sigma}{\sqrt{n}} \quad) = 0.95
\end{aligned}
$$

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

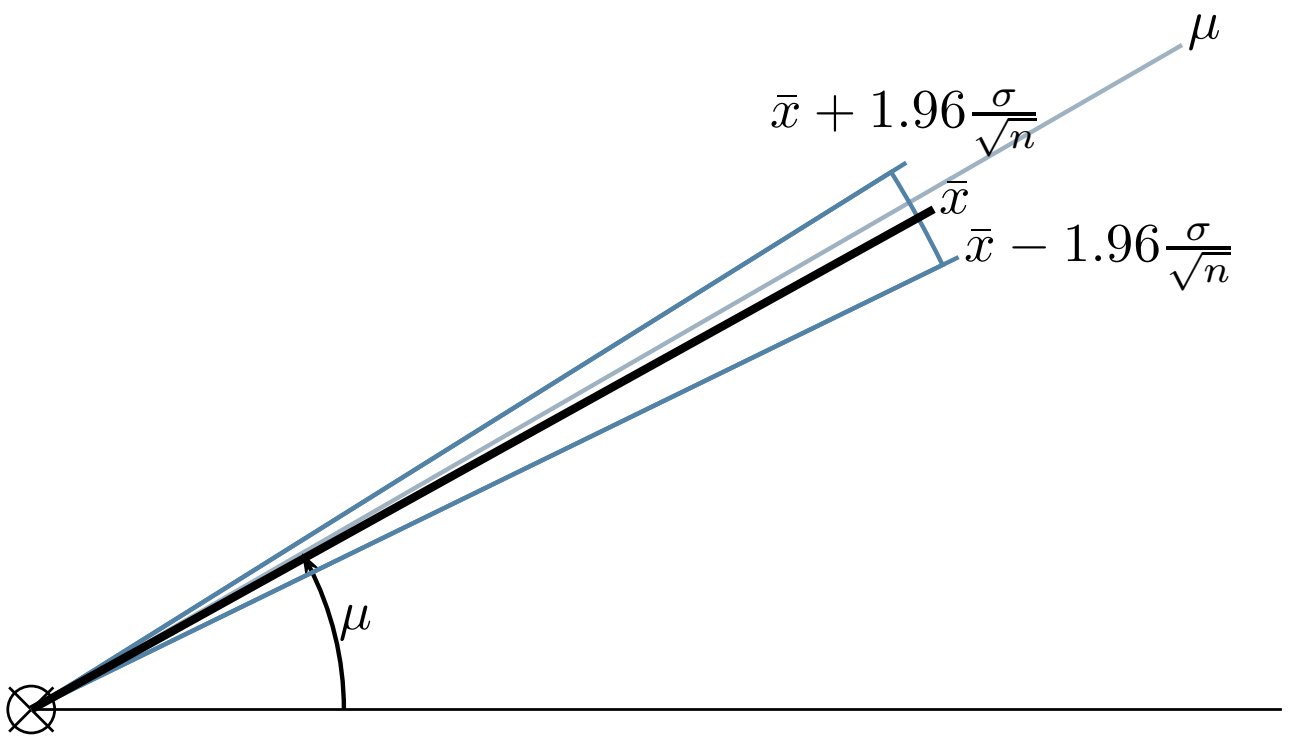
$$\bar{X} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}.$$

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

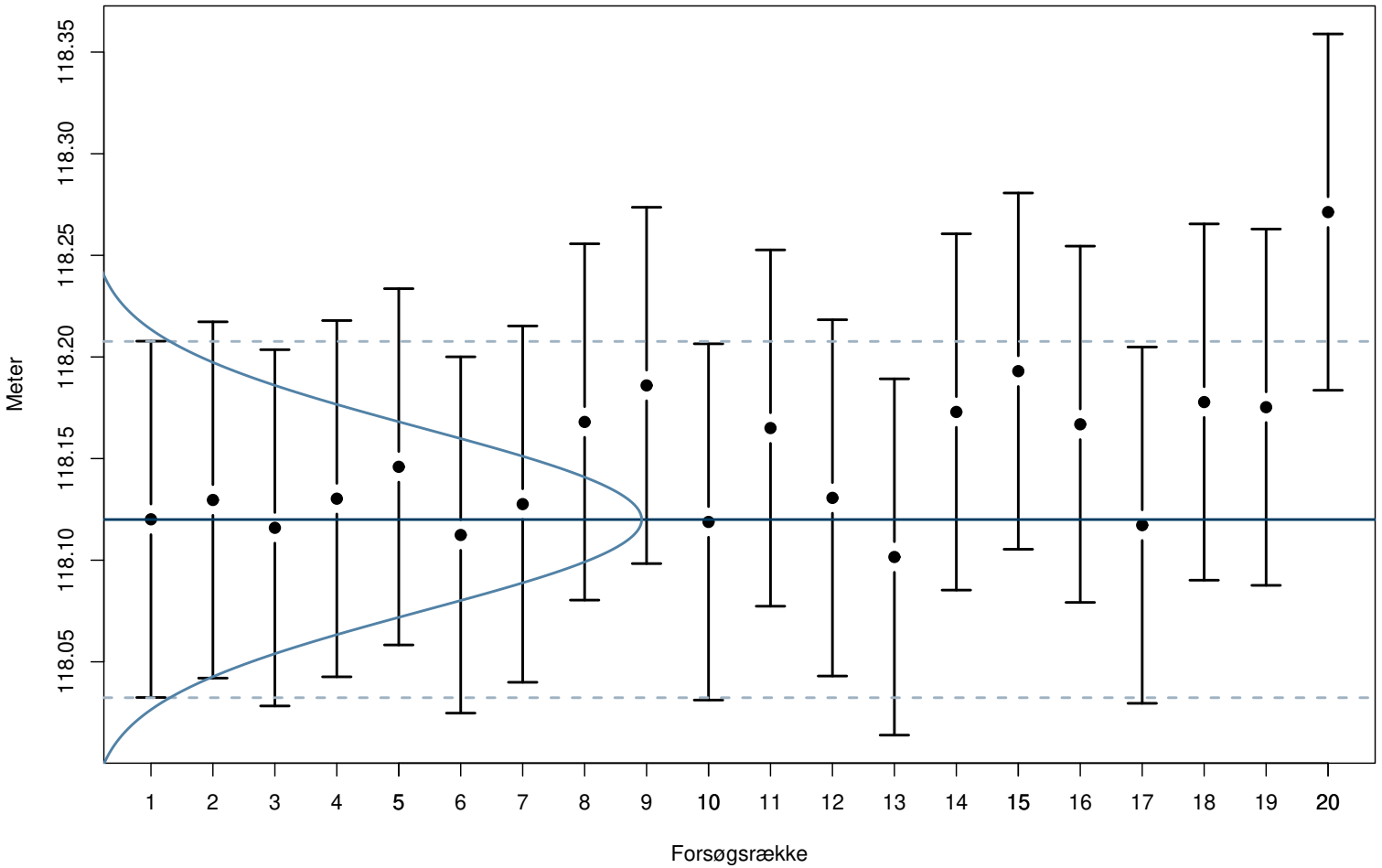
Antag vi observerer de n stokastiske variable k gange, dvs. vi får k observationsrækker med n tal.

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

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

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

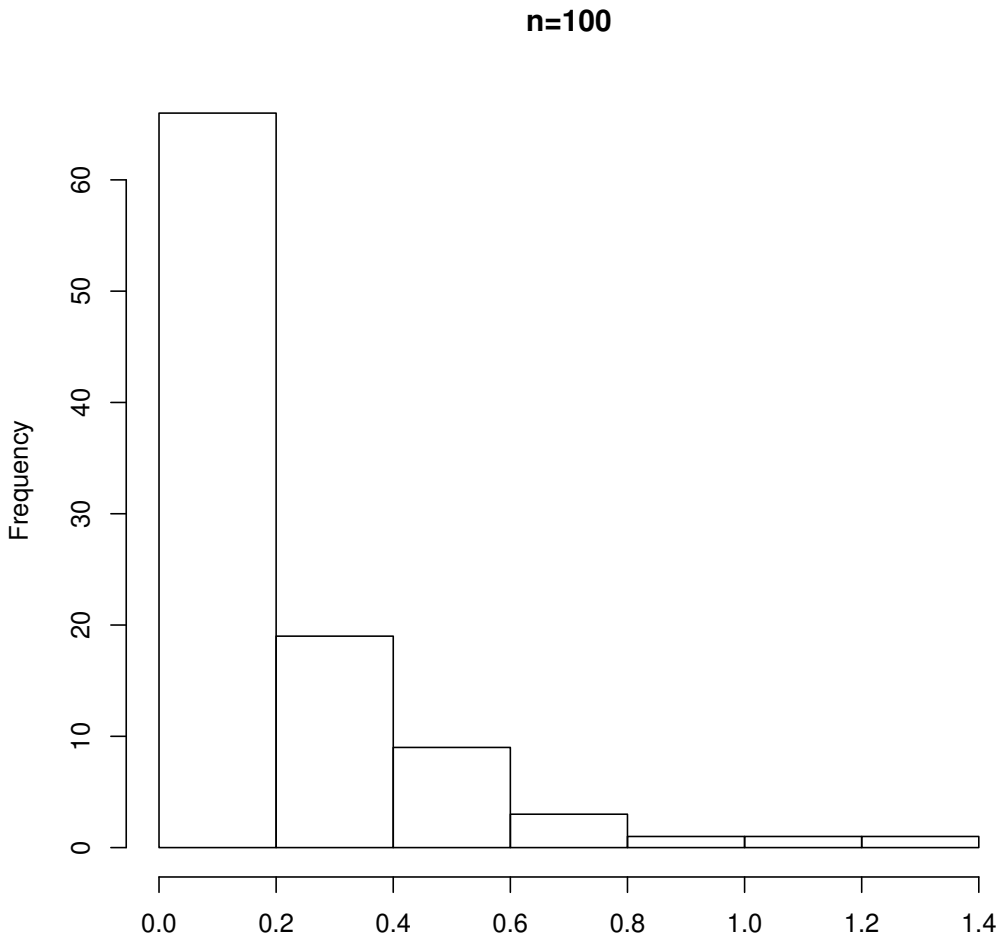
Dvs. selv for ikke-normalfordelte målinger vil konfidensintervallet stadig give god mening.

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.

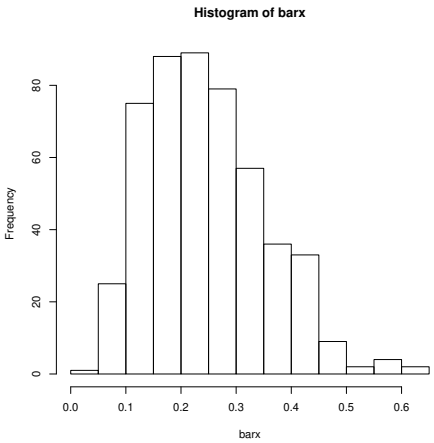
Histogram af første stikprøve $n = 100$



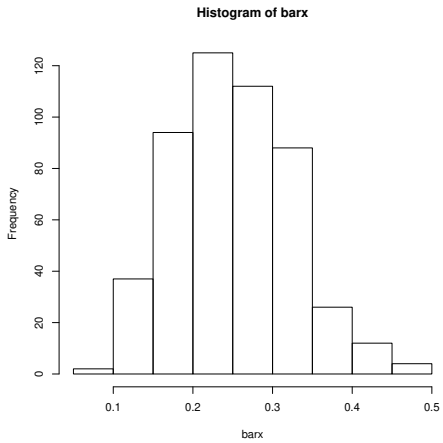
Observationer langt fra normalfordelte !

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

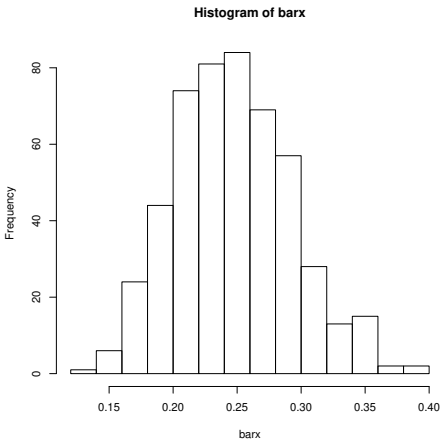
$n = 5$



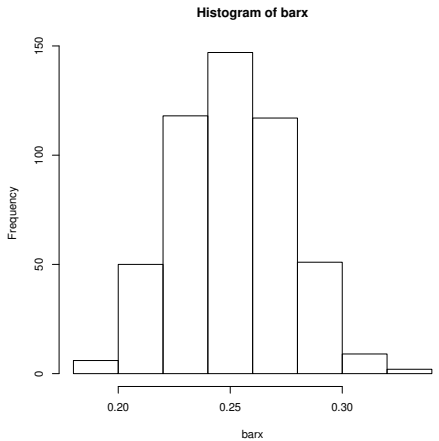
$n = 10$



$n = 30$



$n = 100$



Gennemsnit af mange observationer er normalfordelt !

Estimation af varians

I nogle tilfælde er variansen σ^2 ukendt. Da må vi estimere σ^2 ud fra data.

Som estimator for σ^2 anvendes S^2:

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

Dette estimat er også centralt, dvs.

$$\mathbb{E}(S^2) = \sigma^2$$

Bemærk: S^2 er "empirisk" version af $\mathbb{E}(X-\mu)^2$

Estimater

Har vi observeret data kan vi estimere μ og σ^2 med $\bar{x}$ og s^2. Her udskiftes de stokastiske 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,

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

Eksempel - fortsat

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

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

$$\sum_{i=1}^{10} x_i^2 = 164.508^2 + 164.509^2 + \cdots + 164.514^2 + 164.513^2 = 270638.7 \text{ gon}^2$$

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

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

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

Approksimativt 95% konfidensinterval (erstatter ukendt σ med s):

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} = [164.509; 164.513]$$

Matlab: Stikprøvegennemsnit og -varians

Data:

```
>> x = [164.508,164.509,164.511,164.507,
        164.510,164.511,164.517,164.510,164.514,164.513];
```

Beregn stikprøvegennemsnit $\bar{x}$:

```
>> mean(x)
ans =
  164.5110
```

Beregn stikprøvevariansen s^2:

```
>> var(x)
ans =
   8.8889e-06
```

Beregn spredningen s:

```
>> std(x)
ans =
   0.0030
```

Bonus: Beregn gennemsnit af observation 4 til 7:

```
>> mean(x(4:7)) ans = 164.5113
```

Estimatorer - kendt middelværdi μ

I situationer hvor vi kender μ (fx. på en øvelsesbane hvor sande længder og vinkler er kendt) bruger vi estimatet $\hat{s}$ 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:

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

Bemærk at $\mathbb{V}\mathrm{ar}(\hat{S}^2) \leq \mathbb{V}\mathrm{ar}(S^2)$, dvs. $\hat{s}^2$ er et mere nøjagtigt estimat end s^2.

(gavnligt at bruge al den viden, der er til rådighed)

Vilkårlig transformation af X

Lad X være en stokastisk variabel med

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

samt tæthedsfunktion $f(x)$.

$Y = g(X)$: vilkårlig **differentiabel transformation** af X.

Middelværdien af $Y = g(X)$:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Problem: Middelværdien $\mathbb{E}(Y)$ er ofte vanskelig at beregne.
Løsning: Vi lineariserer transformationen $g(X)$.

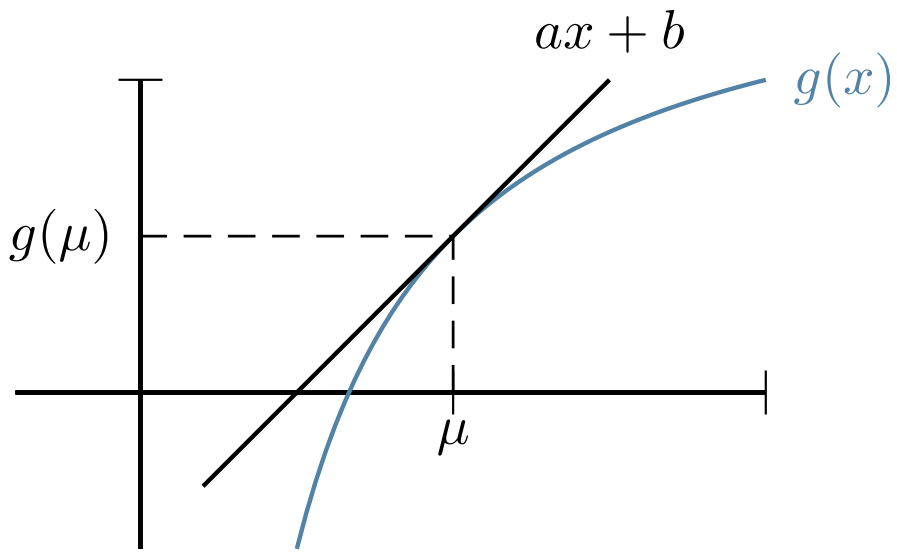
Eksempel på transformationer i landmåling: trigonometriske funktioner, afstandsformel.

Linearisering (approksimering med tangent)

Lineær approksimation af g omkring μ:

$$\begin{aligned} Y = g(X) &\approx g(\mu) + g'(\mu)(X - \mu) \\ &= g'(\mu)X - g'(\mu)\mu + g(\mu) \\ &= aX + b, \end{aligned}$$

hvor $a = g'(\mu)$ og $b = -g'(\mu)\mu + g(\mu)$.



Linearisering

Vi har en approksimation af $g(x)$:

$$Y \approx aX + b,$$

hvor $a = g'(\mu)$ og $b = -g'(\mu)\mu + g(\mu)$.

Heraf følger approksimativ middelværdi og varians for Y:

$$\begin{aligned} \mathbb{E}(Y) &\approx a\mathbb{E}(X) + b \\ &= g'(\mu)\mu - g'(\mu)\mu + g(\mu) \\ &= g(\mu) \end{aligned}$$

$$\begin{aligned} \mathbb{V}\mathrm{ar}(Y) &\approx a^2\mathbb{V}\mathrm{ar}(X) \\ &= g'(\mu)^2\sigma^2, \end{aligned}$$

hvor approximationerne er gode, hvis σ er lille.

Linearisering: Eksempel

Antag $X \sim \mathcal{N}(\mu,\sigma^2)$ og $Y = g(X) = \exp(X)$.

En linearisering af $\exp(x)$ omkring $x = \mu$ giver:

- $g'(x) = \frac{d}{dx}\exp(x) = \exp(x)$
- $g(x) \approx g(\mu) + g'(\mu)(x - \mu) = \exp(\mu) + \exp(\mu)(x - \mu)$.

Heraf følger:

- $\mathbb{E}(Y) \approx g(\mu) = \exp(\mu)$
- $\mathbb{V}\mathrm{ar}(Y) \approx g'(\mu)^2\sigma^2 = (\exp(\mu))^2\,\sigma^2 = \exp(2\mu)\sigma^2$

Vi har derfor, at Y er tilnærmet normalfordelt, med middelværdi $\exp(\mu)$ og varians $\exp(\mu)\sigma^2$:

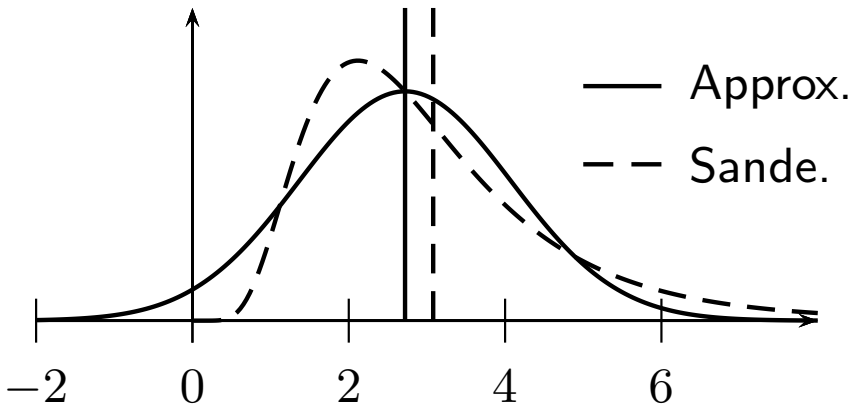
$$Y \approx \mathcal{N}(\exp(\mu), \exp(2\mu)\sigma^2).$$

Linearisering: Eksempel (forts.)

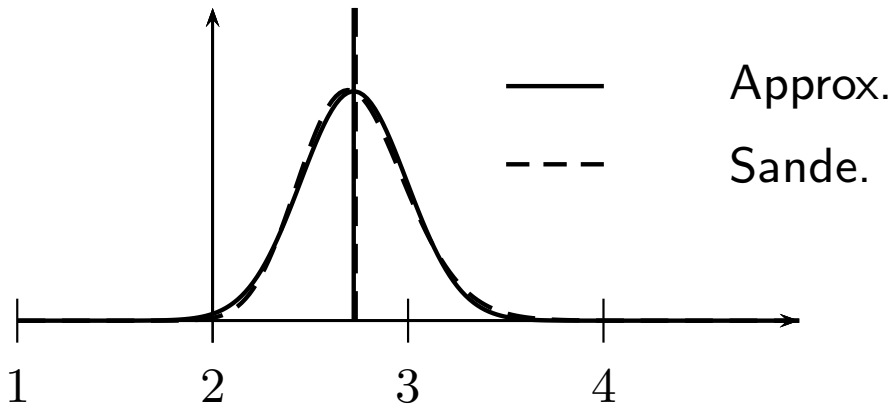
Antag (igen) $X \sim \mathcal{N}(\mu,\sigma^2)$ og $Y = g(X) = \exp(X)$.

To eksempler, hvor $\mu = 1$, og $\sigma = 0.5$ (venstre) og $\sigma = 0.1$ (højre).

$X \sim \mathcal{N}(1; 0{,}5^2)$
$Y \approx \mathcal{N}(\exp(1); \exp(2)0{,}5^2)$



$X \sim \mathcal{N}(1; 0{,}1^2)$
$Y \approx \mathcal{N}(\exp(1); \exp(2)0{,}1^2)$



Til venstre er den relative varians for X $\sigma^2/\mu^2 = 0{,}25$ og til højre er den relative varians for X $0{,}01$. Jo mindre relativ varians jo bedre er approksimationen.

Linearisering — estimation af transformeret størrelse

Antag vi vil estimere $\theta = h(\mu)$ hvor vi kan estimere μ vha. $\bar{X}$ baseret på en stikprøve $X_1, \ldots, X_n$ af uafhængige stokastiske variable med middelværdi μ og varians σ^2.

Da er vores estimat

$$\hat{\theta} = h(\bar{X})$$

Pr. linearisering og central grænseværdisætning har vi

$$\hat{\theta} \approx N(\theta, (h'(\mu))^2 \sigma^2 / n)$$

Dermed er et approksimativ 95% konfidensinterval givet ved

$$\hat{\theta} \pm 1.96 |h'(\bar{X})| \sigma / \sqrt{n}$$

I praksis er σ^2 ofte ukendt og erstattes af estimatet s^2 baseret på $X_1, \ldots, X_n$.

Trigonometriske funktioner: Gon og radianer

Lad $\sin_r(x)$ og $\sin(x)$ betegne sinus når vinklen x er målt i hhv. radianer og gon. Tilsvarende for cosinus og tangens.

Vi har

$$\begin{aligned}\sin(C) &= \sin_r\left(\frac{2\pi}{400\text{ gon}}C\right)\\ &= \sin_r\left(\frac{1}{\omega}C\right),\end{aligned}$$

hvor

$$\omega = \frac{200\text{ gon}}{\pi},$$

er en konverterings-faktor.

Trigonometriske funktioner: Differentitation

Vi har regneregler for differentiation af trigonmetriske funktioner, når vinklen er målt i radianer. Fx.

$$\frac{\mathrm{d}\sin_r(x)}{\mathrm{d}x} = \cos_r(x).$$

Når vinklen er målt i gon får vi:

$$\frac{\mathrm{d}\sin(x)}{\mathrm{d}x} = \frac{\mathrm{d}\sin_r\left(\frac{1}{\omega}x\right)}{\mathrm{d}x} = \cos_r\left(\frac{1}{\omega}x\right)\frac{1}{\omega} = \cos(x)\frac{1}{\omega}.$$

Konverterings-faktoren $\frac{1}{\omega} = \pi/200$ gon optræder på samme måde ved differentiation af cosinus og tangens:

$$\frac{\mathrm{d}\cos(x)}{\mathrm{d}x} = -\sin(x)\frac{1}{\omega} \quad \text{og} \quad \frac{\mathrm{d}\tan(x)}{\mathrm{d}x} = \left(1+\tan(x)^2\right)\frac{1}{\omega}$$