Landmålingens fejlteori
Lektion 2
Repetition
Normalfordelingen
Sum af normalfordelte stokastiske variable
Stikprøve og estimat af middelværdi

Rasmus Waagepetersen - rw@math.aau.dk

Institut for Matematiske Fag
Aalborg Universitet

Repetition - Tæthedsfunktion

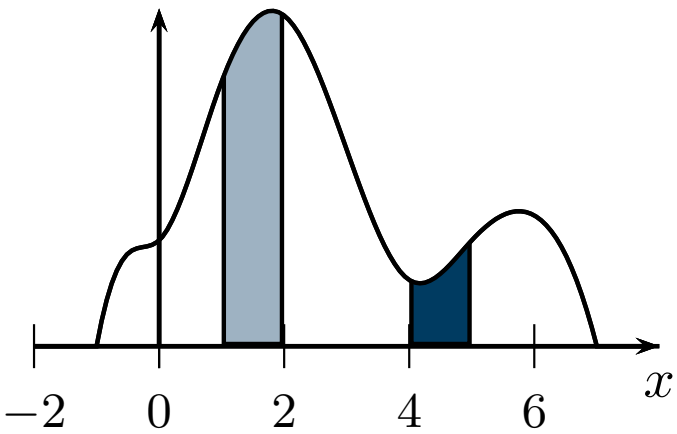
Hvis X er en SV med tæthedsfunktion $f(x)$, så er sandsynligheden for at X ligger mellem a og b ($a \leq b$) givet ved

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x$$

Sandsynligheden for at X ligger mellem a og b er således arealet under $f(x)$ fra a til b.

$P(1 \leq X \leq 2) = \int_1^2 f(x)\mathrm{d}x = \text{Areal}(\square)$

$P(4 \leq X \leq 5) = \int_4^5 f(x)\mathrm{d}x = \text{Areal}(\blacksquare)$



Fordelingsfunktion, middelværdi, varians

Fordelingsfunktion:

$$F(x) = P(X \leq x) \qquad = \int_{-\infty}^{x} f(t)dt$$

Middelværdi (forventet værdi):

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x < \infty$$

Varians:

$$\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) \mathrm{d}x < \infty$$

Spredning/standardafvigelse:

$$\sigma = \sqrt{\mathbb{V}\mathrm{ar}(X)}$$

Eksempel

Ligelig fordelt stokastisk variabel på $[a,b]$:

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

Middelværdi

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

Varians

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

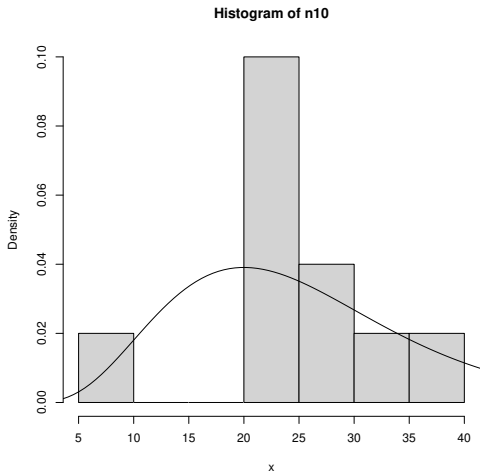
Spredning

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

Tæthedsfunktion og histogram

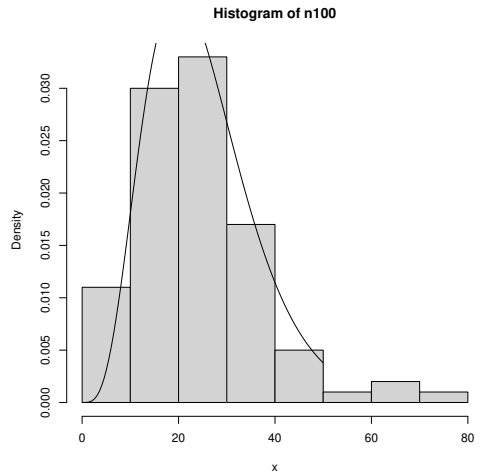
Stikprøver og og tæthedsfunktion for gamma-fordeling med middelværdi $\mu = 25$ og varians $\sigma^2 = 125$:

$n = 10$



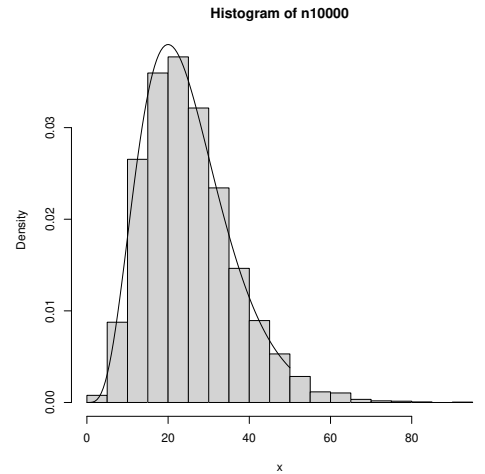
$\bar{X} = 24.60 \quad s^2 = 64.83$

$n = 100$



$\bar{X} = 24.49 \quad s^2 = 167.09$

$n = 10000$



$\bar{X} = 25.21 \quad s^2 = 124.98$

Normalfordelingen

Definition: Normalfordelingen

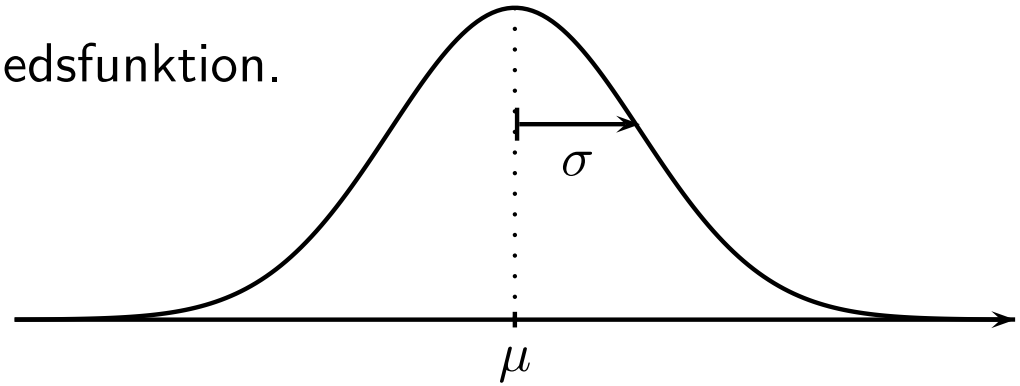
En stokastisk variabel X med **tæthedsfunktion**

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

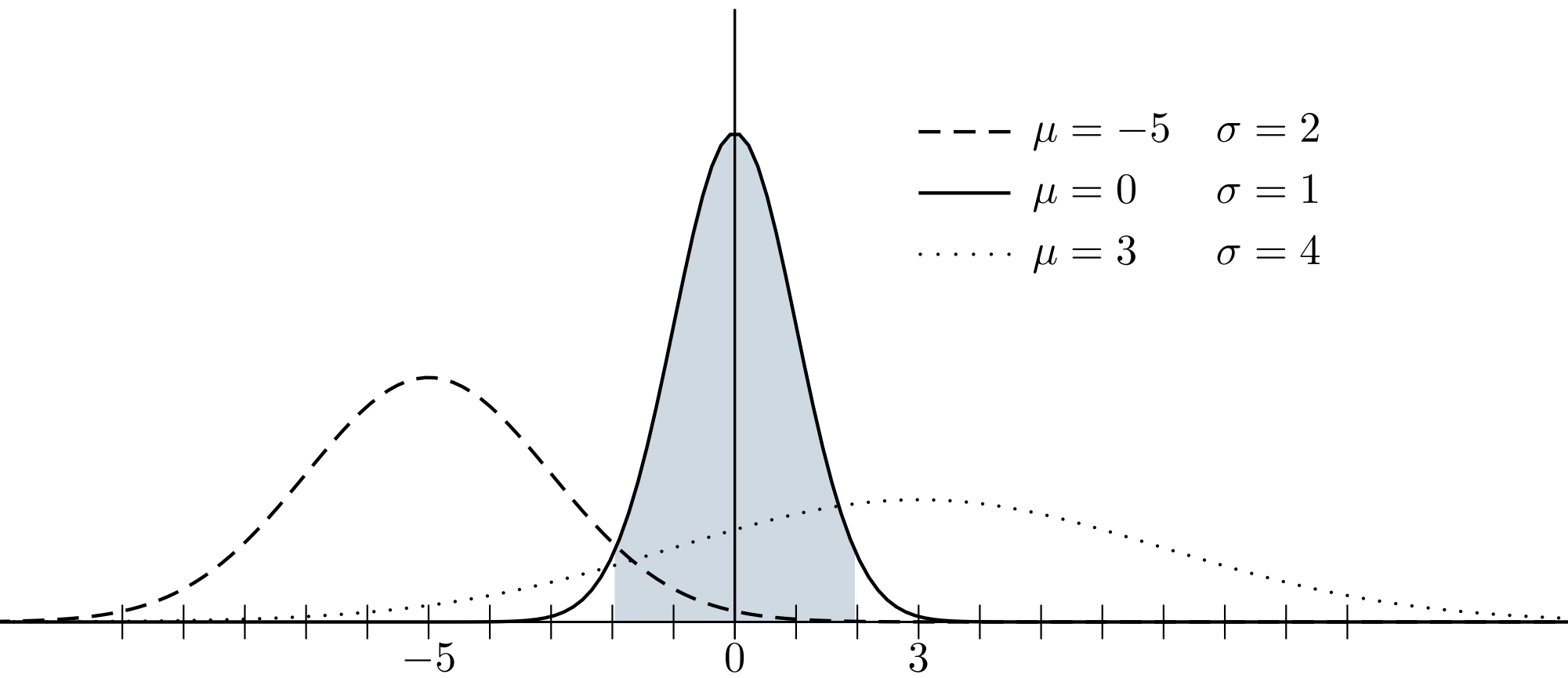
siges at være **normalfordelt** med middelværdi μ og varians σ^2, hvor μ og σ er reelle tal og $\sigma > 0$.

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

Klokkeformet symmetrisk tæthedsfunktion.



Normalfordelingen: Tre eksempler



Fordelingsfunktion for normalfordelingen

Der findes ikke et simpelt 'lukket' udtryk for fordelingsfunktionen.

Sandsynligheder udregnes vha. 'standardisering' og tabelopslag eller vha. `Excel/matlab/python` (senere).

Transformationer

Sætning: Lineær transformation

Antag $X \sim \mathcal{N}(\mu, \sigma^2)$, og a,$b \in \mathbb{R}$ og $a \neq 0$. Lad Y være en lineær transformation af X:

$$Y = aX + b.$$

Da gælder

$$Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

Dvs. en lineær transformation af en normalfordelt SV er stadig normalfordelt.

Middelværdi og varians for Y følger af de generelle regler:

Y har middelværdi: $\mathbb{E}(Y) = \mathbb{E}(aX + b) = a\mathbb{E}(X) + b = a\mu + b$

og varians: $\mathbb{V}\mathrm{ar}(Y) = \mathbb{V}\mathrm{ar}(aX + b) = a^2\mathbb{V}\mathrm{ar}(X) = a^2\sigma^2.$

Tilfældig fejl

Definition: Tilfældig fejl

Antag vi vil måle en længde. Den sande længde betegnes μ. Lad X være en SV, der repræsenterer målingen af længden μ. Vi har

$$X = \mu + \epsilon,$$

hvor ϵ er en *tilfældig fejl*.

Dvs. målingen er den sande længde plus en tilfældig fejl.

Vi antager ofte

$$\epsilon \sim \mathcal{N}(0,\sigma^2).$$

Dvs. den tilfældige fejl er normalfordelt og i middel er fejlen nul.

Det følger, at

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

Standard-normalfordelingen

Definition: Standard-normalfordelingen

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

Tæthedsfunktionen for $\mathcal{N}(0,1)$ er givet ved:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)$$

Den tilhørende fordelingsfunktion betegnes $\Phi(z)$:

Dvs $F(z) = \Phi(z)$ når $\mu = 0$ og $\sigma = 1$.

Der findes ikke et 'lukket' udtryk for fordelingsfunktionen $\Phi(z)$.

Standardisering

Vi kan standardisere en hvilken som helst normalfordelt SV X til at være standard-normalfordelt:

Sætning: Standardisering

Hvis $X \sim \mathcal{N}(\mu, \sigma^2)$, så gælder

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

Middelværdi og varians beregnes vha. foregående sætninger:

$$\mathbb{E}\left(\frac{X-\mu}{\sigma}\right) = \mathbb{E}\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{\mathbb{E}(X)}{\sigma} - \frac{\mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

$$\mathbb{V}\mathrm{ar}\left(\frac{X-\mu}{\sigma}\right) = \mathbb{V}\mathrm{ar}\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = \frac{\mathbb{V}\mathrm{ar}(X)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

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

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Dvs. for X med vilkårlig middelværdi og varians kan vi finde sandsynligheder vha. tabel for standardnormalfordelingen.

Nu om dage benyttes dog typisk i praksis et computer program - f.eks. `matlab`.

Eksempel

Lad $X \sim \mathcal{N}(3,4)$, så er

$$Z = \frac{X-3}{\sqrt{4}} \sim \mathcal{N}(0,1).$$

Udregning af $P(X \leq 5)$ i Matlab:

```
>> normcdf(5,3,sqrt(4))
ans =
    0.8413
```

Vha. standardisering:

```
>> normcdf((5-3)/sqrt(4))
ans =
    0.8413
```

Inverse fordelingsfunktion

Den inverse fordelingsfunktion Φ^{-1} går den "modsatte vej" af Φ. Dvs. hvis $\Phi(1.4) = 0.8849303$, så er $\Phi^{-1}(0.8849303) = 1.4$.

1.4 er 0.8849303 *fraktilen* for standardnormalfordelingen.

Inverse fordelingsfunktion $\Phi^{-1}(p)$ i `matlab`:

```
>> norminv(0.8849303)
ans =
    1.4
```

Eksempel:
Antag $Z \sim \mathcal{N}(0,1)$. Find konstant z, så $P(Z \leq z) = 0{,}8$.

Løsning:

$$P(Z \leq z) = \Phi(z) = 0{,}8 \Leftrightarrow z = \Phi^{-1}(0{,}8) = 0.8416$$

```
>> norminv(0.8)
ans =
    0.8416212
```

Opsummering: sandsynligheder og fraktiler for normalfordeling

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

$p = P(X \leq x) = F(x)$ kan udregnes enten vha. omformning til standardnormalfordeling og tabel eller direkte vha. f.eks. `matlab` (eller Excel eller python eller ...).

For et givet $0 \leq p \leq 1$ kaldes

$$x_p = F^{-1}(p)$$

for p-*fraktilen*. Dvs. x_p opfylder $P(X \leq x_p) = p$.

Kan også udregnes vha. standardnormalfordeling (og baglæns opslag i tabel) eller `matlab`.

Eksempel $X \sim N(-2,3)$.

Beregn $P(X \geq -2.908288)$ samt $x_{0.3}$:

```
>> 1-normcdf(-2.908288,-2,sqrt(3))
ans =
    0.7
>> norminv(0.3,-2,sqrt(3))
ans =
    -2.908288
```

Eksempler

Hvad hvis vi gerne vil udregne $P(-3 \leq X \leq 0)$?

Hvordan finder vi x så $P(X \geq x) = 0.1$?

Uafhængige stokastiske variable

To stokastiske variable X_1 og X_2 er 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 af fejl ϵ_1 er for stor ville det være uheldigt, hvis det medførte en forøget sandsynlighed for, at fejlen ϵ_2 også bliver for stor.

Uafhængighed fortsat

Den betingede sandsynlighed kan udregnes som

$$P(X_2 \leq x_2 \,|\, X_1 \in [a,b]) = \frac{P(X_2 \leq x_2 \text{ og } X_1 \in [a,b])}{P(X_1 \in [a,b])}$$

Dermed er X_1 og X_2 uafhængige hvis

$$P(X_1 \in [a,b] \text{ og } X_2 \in [c,d]) = P(X_1 \in [a,b])P(X_2 \in [c,d])$$

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 $a_0, a_1, \ldots, a_n \in \mathbb{R}$ er reelle tal, så gælder

$$\begin{aligned}\mathbb{E}(a_0 + a_1X_1 + \cdots + a_nX_n) &= a_0 + a_1\mathbb{E}(X_1) + \cdots + a_n\mathbb{E}(X_n) \\ &= a_0 + a_1\mu_1 + \cdots + a_n\mu_n\end{aligned}$$

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

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

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

Sum af normalfordelte SV

Sætning: Sum af normalfordelte SV

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

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

og $a_0, \ldots, a_n$ 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 $X_1, X_2, \ldots X_n$ er uafhængige er variansen af summen

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

Eksempel

X har middelværdi 4 og spredning 1, Y middelværdi -1 og spredning 3, Z har middelværdi 0 og spredning 2.

Eksempel

X har middelværdi 4 og spredning 1, Y middelværdi -1 og spredning 3, Z har middelværdi 0 og spredning 2.

Hvad er middelværdi og spredning af $2X - Y + 3Z$?

Middelværdi $2 \cdot 4 + (-1) \cdot (-1) + 3 \cdot 0 = 9$.

Hvis X,Y,Z uafhængige: varians $2^2 \cdot 1^2 + (-1)^2 \cdot 3^2 + 3^2 \cdot 2^2 = 49$

Spredning $7 = \sqrt{49}$ (bemærk: $\neq 2 \cdot 1 + (-1) \cdot 3 + 3 \cdot 2 = 5$)

"spredning af sum er ikke lig sum af spredninger"

Hvis X,Y,Z er normalfordelte er $2X - Y + 3Z \sim \mathcal{N}(9,49)$

Sandsynligheds-intervaller for normalfordeling

Lad $X \sim \mathcal{N}(\mu,\sigma^2)$ og $k = 1,1.96,2,3$. Interval omkring μ:
$[\mu - k\sigma; \mu + k\sigma]$

Sandsynlighed for at X ligger i interval:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq \frac{X-\mu}{\sigma} \leq k)$$

$$= P(-k \leq Z \leq k) = \begin{cases} 0.68 & k = 1 \\ 0.95 & k = 1.96 \\ 0.954 & k = 2 \\ 0.997 & k = 3 \end{cases}$$

Bemærk: sandsynlighed afhænger kun af k - ikke af de konkrete værdier af μ og σ.

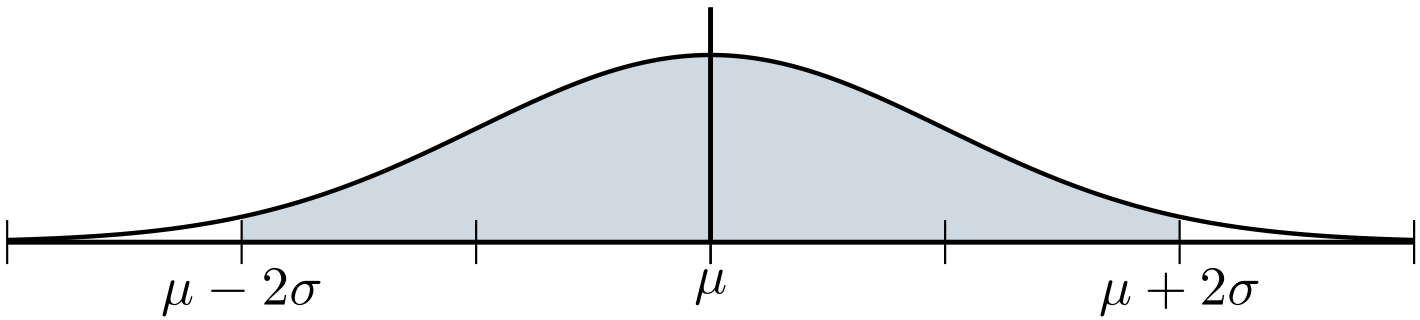
Vi kan oversætte 'antal σ' til sandsynligheder. Dvs. nyttigt at måle 'bredde' af normalfordeling i antal σ.

Statistisk model

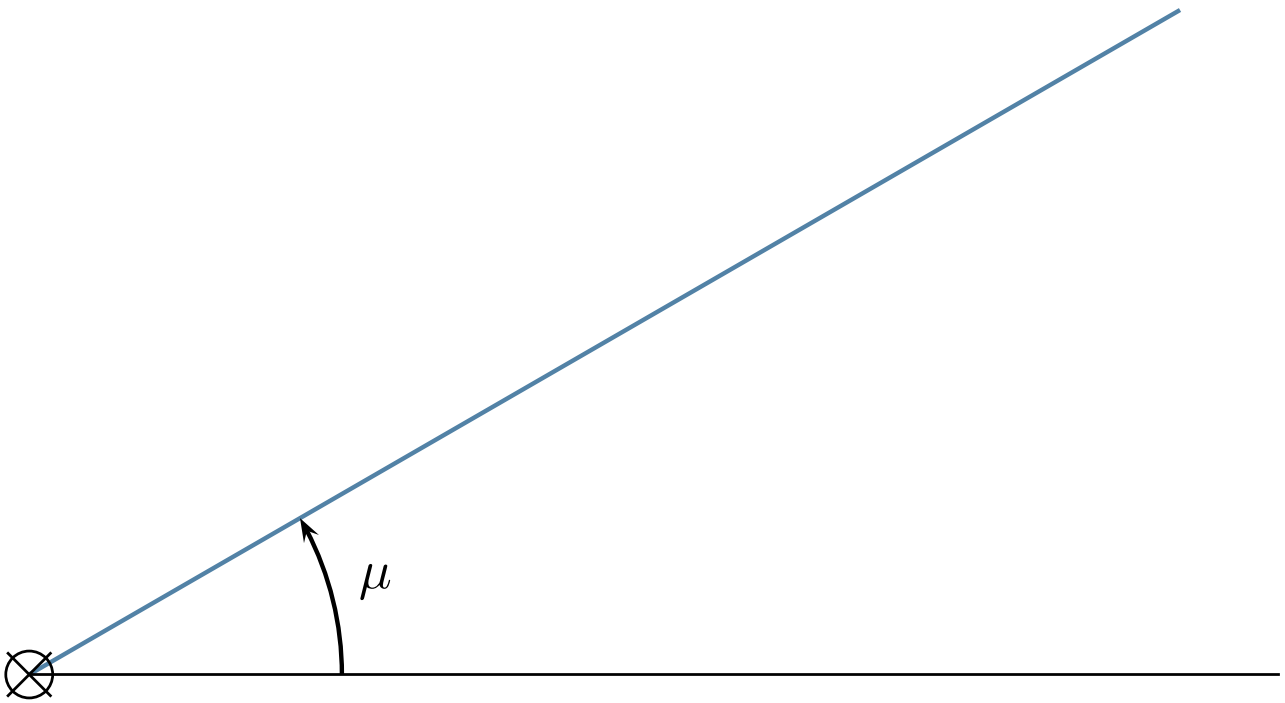
Model for vinkler i landmåling:

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),

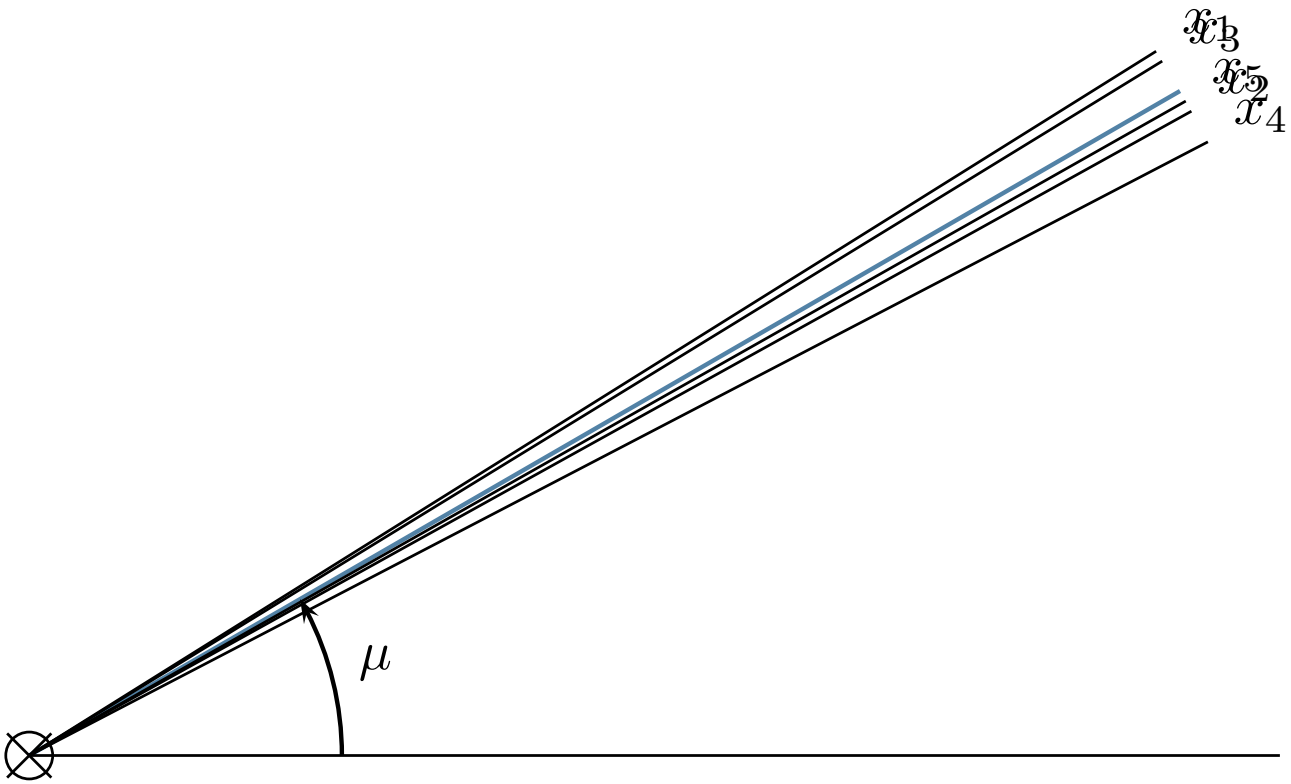
$$X_i \sim \mathcal{N}(\mu,\sigma^2), \quad i = 1,\ldots,n.$$



Model for vinkler



Model for vinkler



Observationer

Ved opmåling af en vinkel foretages n **observationer** $x_1,\ldots,x_n$ som er **realisationer** af de stokastiske 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 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

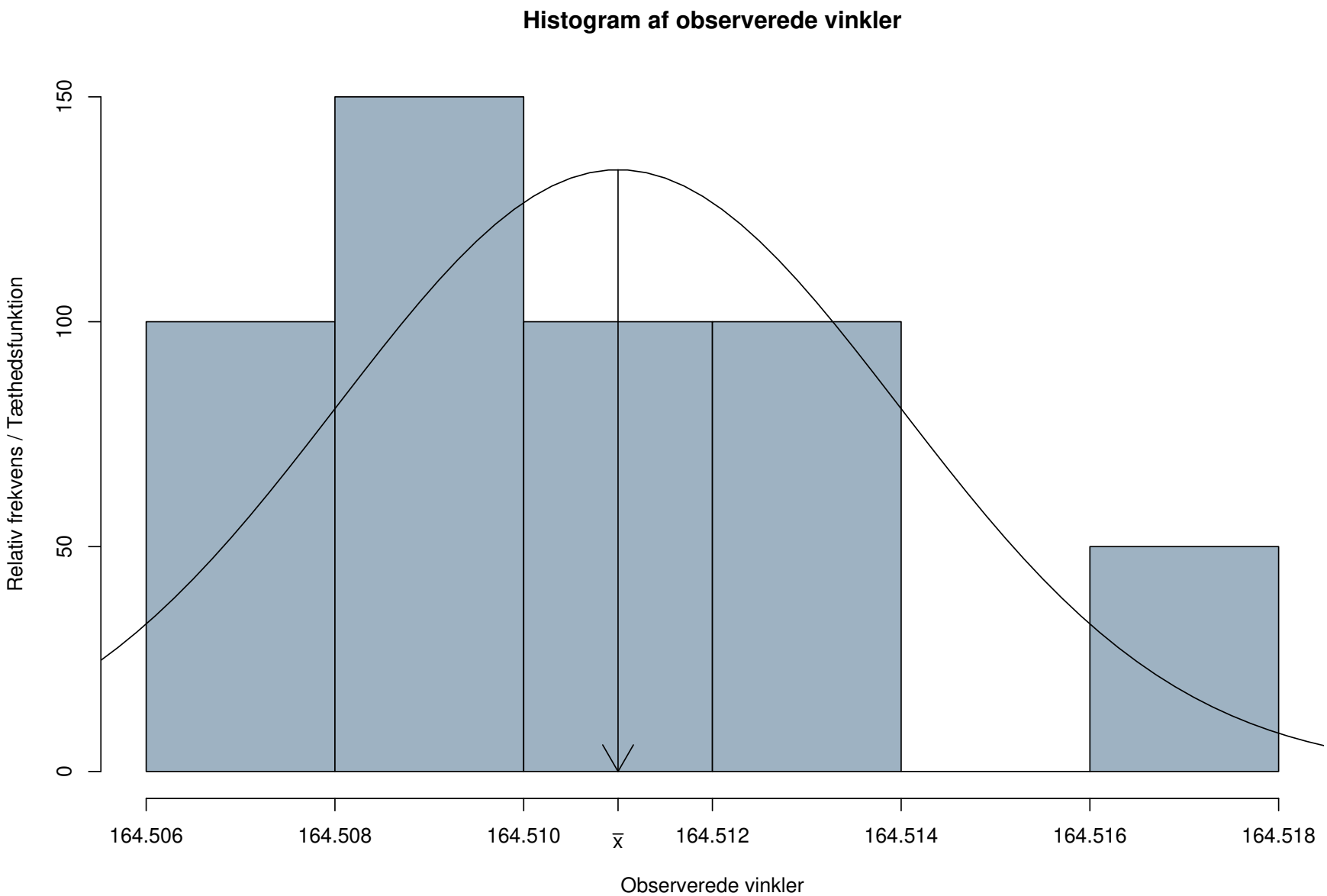
Jf. eksempel fra noterne observeres følgende 10 satser af en vinkel.

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 $\bar{X}$,

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

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

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

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

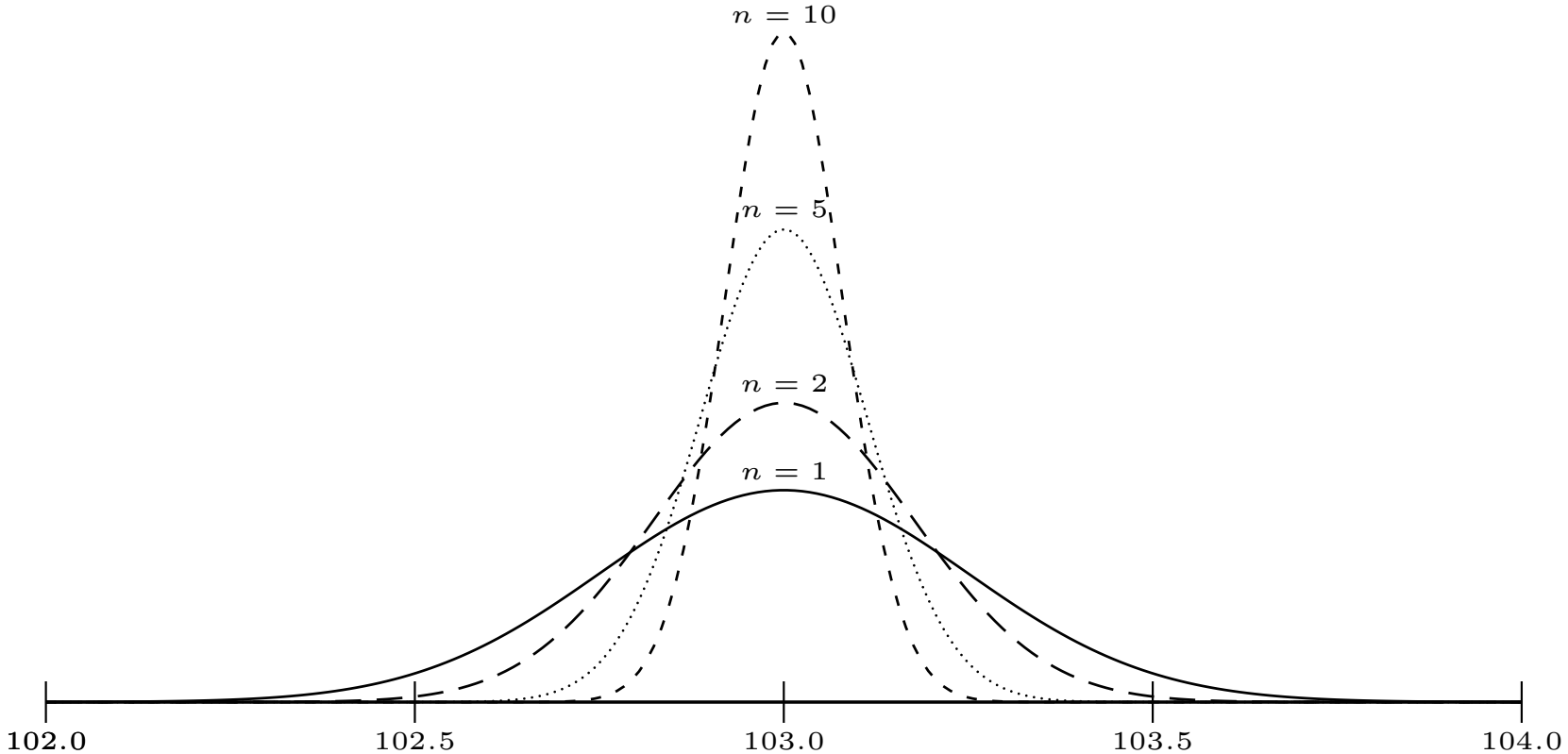
$$\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = n\frac{1}{n^2}\sigma^2 = \frac{1}{n}\sigma^2$$

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.

Effekt af at øge n

Hvis vi har $n = 100$ bliver varians 100 gange mindre.

Imidlertid er spredningen $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ mere nyttig til at vurdere usikkerhed.

Dvs. spredningen bliver kun $10 = \sqrt{100}$ gange mindre.

Altså bør man se på $\sqrt{n}$ for at få et retvisende indtryk af effekten på usikkerheden af at bruge n gentagelser.