Landmålingens fejlteori
Lektion 1
Det matematiske fundament
Kontinuerte stokastiske variable

Rasmus Waagepetersen - rw@math.aau.dk

Institut for Matematiske Fag
Aalborg Universitet

Landmålingens fejlteori - lidt om kurset

Kursusholder
Rasmus Waagepetersen
Institut for Matematiske Fag
`rw@math.aau.dk`

Litteratur
Kasper Klitgaard Berthelsen, Poul Winding & Jens Møller Pedersen, *Noter i Fejlteori.*

Kursusform
Seks kursusgange, hver med vekslen mellem forelæsninger og opgaver.

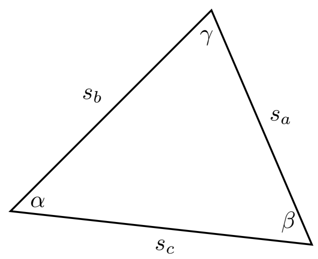
Eksamen
Mundtlig og fælles med *Det matematiske grundlag for kortprojektioner*.
Tager udgangspunkt i seks opgaver - tre fra hvert af de to kurser.

Formål

- introduktion til basal sandsynlighedsberegning
- introduktion til basal statistik/fejlteori
- hvordan kvantificeres usikkerhed på en måling eller et gennemsnit af målinger ?
- hvorledes kvantificeres usikkerhed på en størrelse beregnet ud fra målinger som er behæftet med fejl ?

NB: jeg afviger af og til lidt fra noternes fremstilling, når det er hensigtsmæssigt.

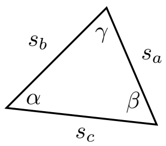
Formål



Arealet af ovenstående trekant kan bestemmes vha:
areal $= \frac{1}{2} s_b \cdot s_c \cdot \sin \alpha$.

Hvordan kvantificerer vi usikkerheden på målinger/skøn af α, s_b og s_c ?
Hvordan kvantificerer vi den resulterende usikkerhed på det beregnede areal ?

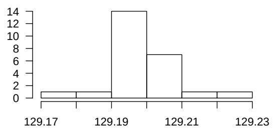
Formål



Antag vi har følgende 25 målinger af α:

129.188, 129.203, 129.211, ..., 129.196, 129.205, 129.193.

Histogram over målingerne:



Hvad er et godt skøn/estimat for α? Gennemsnittet? Hvad kan vi sige om usikkerheden på vores estimat?

Kender vi usikkerheden på skønnene/estimaterne af α, s_b og s_c giver Fejlforplantningsloven (clou'et i dette kursus) et skøn over usikkerheden på det beregnede areal.

Sandsynlighed

Udgangspunkt: Vi udfører et eksperiment/måling og observerer om en given hændelse H indtræffer.

Antag vi gentager eksperimentet n gange under identiske forhold og lader m angive antal gange, hvor H indtraf.

Hvis n går mod uendelig vil andelen m/n gå mod en fast størrelse - sandsynligheden for at H indtræffer.

Sandsynligheden noteres $P(H)$ ('P' for 'probability')

Eksempel: Kast med en fair mønt

Eksperiment: Kast en *fair* mønt

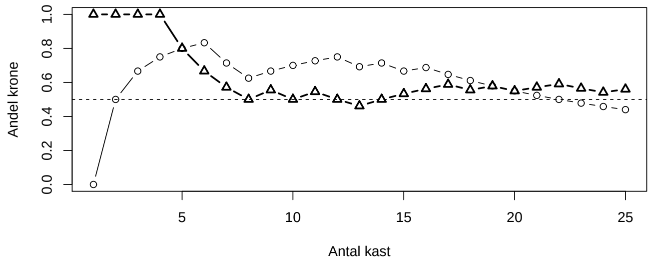
Mulige udfald: *Plat* og *Krone*

H: kast giver krone.

Mønten er fair, dvs. i det lange løb forventer vi lige store andele plat og krone. Med andre ord gælder $P(H) = 0{,}5$.

Eksempel - fortsat

Vi har to gange kastet en fair mønt 25 gange.
Resultat: to følger af plat (P) og krone (K)
1. følge: P K K K K K P P K K K K P K P K P P P P P P P P P (○)
2. følge: K K K K P P P P K P K P P K K K K P K P K K P P K (△)



Bemærk hvordan den observerede andel nærmer sig 0,5 i begge følger.

Eksempel - måling af længde

Antag vi måler en længde. Eksempel på hændelse: den målte længde er mellem 173,21 og 173,24 meter.

Mål længden igen og igen (under samme forhold og uafhængigt af hinanden). Sandsynligheden

$$P(\text{den målte længde er mellem 173,21 og 173,24})$$

er således andelen af målinger, der (i det lange løb) falder mellem 173,21 og 173,24.

For en generel hændelse H gælder følgende regler

Definition: Hændelser og sandsynligheder

For en hændelse H gælder
1. $0 \leq P(H) \leq 1$
2. $P(\text{ej } H) = 1 - P(H)$.

Sandsynligheden $P(H)$ er stadig andelen af eksperimenter, hvor hændelsen H indtræffer (i det lange løb).

Hændelsen "ej H" kaldes den *komplementære* hændelse til H.

Vi ved med sikkerhed at enten H eller ej H indtræffer. Dvs.
$P(H \text{ eller ej } H) = P(H) + P(\text{ ej } H) = 1$.

Stokastisk variabel

Stokastisk (tilfældig) variabel er den matematiske terminologi for en størrelse (f.eks. en fremtidig måling), der er behæftet med usikkerhed.

En stokastisk variabel (SV) er blot en reel variabel, hvis forskellige udfald tillægges sandsynligheder.

Stokastiske variable noteres ofte med store bogstaver, mens konkrete udfald noteres med tilsvarende små bogstaver.

F.eks. kan X angive en fremtidig (endnu ikke foretaget måling) mens x angiver det konkrete resultat, når målingen er foretaget - f.eks. $x = 2$.

Vi kan da tale om sandsynligheden for hændelser $X = x$, $X \leq x$, $X > 4$, dvs. $P(X = x)$, $P(X \leq x)$, $P(X > 4)$...

Vi har nu brug for en bekvem måde at specificere sandsynlighederne for X !

Undtagelse: for en trekant følger vi traditionen og lader A,B,C og a,b,c betegne henholdsvis vinkler og sidelængder.

Kontinuert og diskret stokastisk variabel

Diskrete stokastiske variable er variable hvor $P(X = x) > 0$ for en mængde af diskrete udfald - f.eks. $x = 0,1,2,3,...$. (terningkast, tælledata,...).

Sandsynligheder for diskrete stokastiske variable specificeres vha. sandsynlighedsfunktion $p(x) = P(X = x)$ for de diskrete udfald x.

Kontinuerte stokastiske variable kan som udgangspunkt antage alle reelle værdier.

Målinger af vinkler og længder betragtes mest relevant som kontinuerte variable.

Sandsynligheder for kontinuerte stokastiske variable

For en kontinuert variabel tillægger vi kun positive sandsynligheder til intervaller, f.eks. $P(1 < X \leq 2)$.

For et enkelt x har vi altid $P(X = x) = 0$ for alle x.

Sandsynligheder for en kontinuert stokastisk variabel specificeres vha. tæthedsfunktion eller fordelingsfunktion.

Hvorfor nok at give sandsynligheder til intervaller ?

Målinger er i praksis altid afrundede størrelser.

Måling 201,43 med to decimaler repræsenterer i realiteten et interval $[201{,}425{,}201{,}435[$ - vi kender ikke den nøjagtige værdi.

Dvs. i realiteten opererer vi altid med (små) intervaller omkring de målte størrelser.

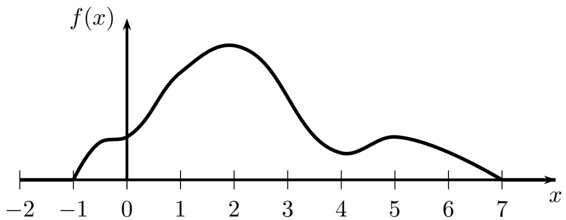
Tæthedsfunktion

Definition: Tæthedsfunktion

En tæthedsfunktion $f(x)$ er en reel funktion, der opfylder

1. $f(x) \geq 0$ for alle $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Eksempel på en tæthedsfunktion:



Hvordan skal tæthedsfunktionen fortolkes?

Tæthedsfunktion: Fortolkning

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

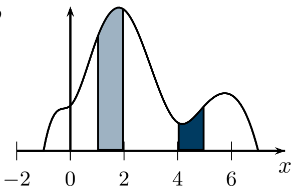
$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x$$

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

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

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

Dvs. $P(1 \leq X \leq 2) > P(4 \leq X \leq 5)$.



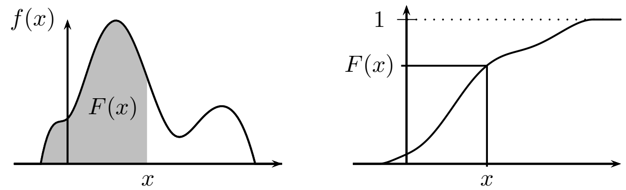
Fordelingsfunktionen

Definition: Fordelingsfunktion

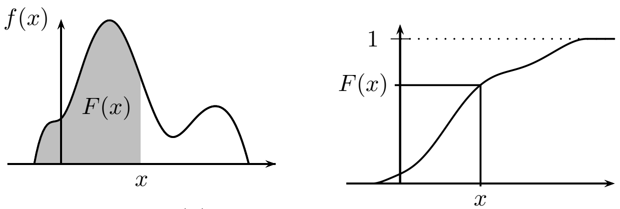
Lad X være en stokastisk variabel med tæthedfunktion $f(x)$. Da er fordelingsfunktionen $F(x)$ givet ved

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

Eksempel på tæthedsfunktion og tilhørende fordelingsfunktion:



Fordelingsfunktionen: Egenskaber



Fordelingsfunktionen $F(x)$ har følgende egenskaber:

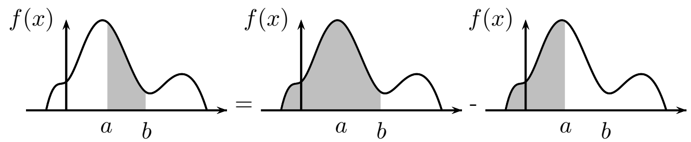
- $F(x)$ er en ikke-aftagende funktion
- $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to\infty} F(x) = 1$
- $F'(x) = f(x)$

Bemærk: entydig sammenhæng mellem f og F !

Fordelingsfunktionen: Egenskaber

For en SV med fordelingsfunktion $F(x)$ har vi

$$P(a \leq X \leq b) = F(b) - F(a).$$



Sandsynligheden for at X tager værdien a er nul:

$$P(X = a) = P(a \leq X \leq a) = F(a) - F(a) = 0$$

"Mindre end" og "Mindre end eller lig" er derfor det samme:

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$$

Middelværdi - frekventiel definition

Fortolkning: Lad $X_1, X_2, \ldots$ repræsentere gentagne målinger af samme størrelse under ens betingelser.

Så svarer den fælles middelværdi $\mu = \mathbb{E}X_1 = \mathbb{E}X_2 = \ldots$ til grænseværdien af det empiriske gennemsnit:

$$\frac{1}{n}(x_1 + x_2 + \cdots + x_n) \to \mu \quad \text{når } n \to \infty$$

hvor $x_1, x_2, \ldots$ er udfald af $X_1, X_2, \ldots$.

Middelværdi - definition ud fra tæthed

Definition: Middelværdi

Lad X være en kontinuert stokastisk variabel med tæthedsfunktion f. **Middelværdien** af X defineres som

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

- også kaldet forventet værdi (UK: $\mathbb{E}$xpectation).

Bemærk: $f(x)\mathrm{d}x \approx P(x \leq X \leq x + \mathrm{d}x)$.

Integralet kan opfattes som et vægtet gennemsnit af X's udfald x med vægte $f(x)$ givet ved X's tæthedsfunktion.

$$\int_{-\infty}^{\infty} x f(x) \mathrm{d}x \approx \sum_{i=1}^{k} x_i f(x_i)\Delta \approx \sum_{i=1}^{k} x_i P(X \in]x_i, x_i + \Delta])$$

Middelværdi og transformation af SV

Sætning: Middelværdi af generel transformation

Antag X er en SV med tæthedsfunktion $f(x)$. Lad $Y = h(X)$. Da er Y er SV med middelværdi

$$\mathbb{E}[Y] = \mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) \mathrm{d}x$$

Typisk er integralet svært at udregne pånær når $h(x)$ er en lineær funktion af x.

Middelværdi af lineær transformation

Sætning: Middelværdi af lineær transformation

Antag X er en SV med tæthedsfunktion $f(x)$ og middelværdi μ. Lad $Y = aX + b$. Da er Y en SV med middelværdi

$$\mathbb{E}[Y] = \mathbb{E}[aX + b] = a\mathbb{E}[X] + b = a\mu + b$$

Bevis:

$$\begin{aligned}\mathbb{E}[aX + b] &= \int_{-\infty}^{\infty} (ax + b) f(x) \mathrm{d}x \\ &= a \int_{-\infty}^{\infty} x f(x) \mathrm{d}x + b \int_{-\infty}^{\infty} f(x) \mathrm{d}x \\ &= a\mathbb{E}[X] + b = a\mu + b.\end{aligned}$$

Eksempel: Telefonabonnement

Den månedlige samtaletid for en tilfældig abonnent hos TeleFlunk er i middel 83,2 minutter. Prisen pr. måned er 99kr/md + 0,10 kr/min.

Hvad er den månedlige udgift i middel?

Lad

- Lad X være den (tilfældige) månedlige samtaletid målt i minutter.
- Lad Y være den månedlige udgift, dvs. $Y = 99 + 0{,}1X$.

Middel-udgiften er

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(99 + 0{,}1X)\\ &= 99 + 0{,}1\mathbb{E}(X)\\ &= 99 + 0{,}1 \cdot 83{,}2\\ &= 107{,}32.\end{aligned}$$

Middel-udgiften er altså 107,32 kr.

Varians og spredning

Definition: Varians

Lad X være en kontinuert stokastisk variabel med tæthedsfunktion f. **Variansen** af X defineres som

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

hvor σ kaldes **spredningen** (eller **standardafvigelsen**) - mål for forventet variabilitet i X.

Variansen er den *forventede værdi af kvadratet på afvigelsen fra middelværdi* - dvs. hvis X ofte langt fra μ så er σ^2 stor !

Spredning/standardafvigelse:

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

Nyttig omskrivning af defintion for varians:

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

Varians af en lineær transformation

Sætning: Varians af lineær transformation

Antag X er en SV med tæthedsfunktion $f(x)$, middelværdi μ og varians σ^2. Lad $Y = aX + b$. Da er Y en SV med varians

$$\mathbb{V}\mathrm{ar}[Y] = \mathbb{V}\mathrm{ar}[aX+b] = a^2\mathbb{V}\mathrm{ar}[X] = a^2\sigma^2.$$

Bevis:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[Y] &= \mathbb{E}[(Y-\mathbb{E}[Y])^2]\\ &= \mathbb{E}[(aX+b-(a\mu+b))^2]\\ &= \mathbb{E}[(a(X-\mu)^2)]\\ &= a^2\mathbb{E}[(X-\mu)^2] = a^2\mathbb{V}\mathrm{ar}[X] = a^2\sigma^2\end{aligned}$$

Bemærk at b *ikke* indgår.

Eksempel: Telefonabonnement

Standardafvigelsen for den månedlige samtaletid for en tilfældig abonnent hos TeleFlunk er 17,2 minutter.

Hvad er variansen for den månedlige udgift?

$$\begin{aligned}\mathbb{V}\mathrm{ar}(Y) &= \mathbb{V}\mathrm{ar}(99 + 0{,}1X)\\ &= 0{,}1^2\mathbb{V}\mathrm{ar}(X)\\ &= 0{,}1^2 \cdot 17{,}2^2\\ &= 2{,}96.\end{aligned}$$

Variansen er altså 2,96 kr^2.

Standardafvigelsen for udgiften er $\sqrt{2{,}96 \text{ kr}^2} = 1{,}72$ kr.

Varians vs. standardafvigelse

Fordel ved standardafvigelse: samme enhed for den stokastiske variabel og dens standardafvigelse (jf. foregående eksempel).

Standardafvigelse bekvemt mål for usikkerhed - f.eks. kan $\pm 2\sigma$ fortolkes som et 95% interval for en stokastisk variabel (vender tilbage til dette senere).

Normalfordelingen

Definition: Normalfordelingen

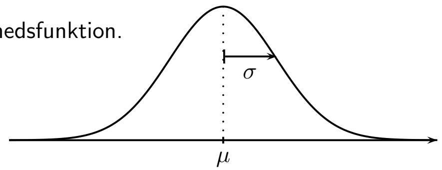
En stokastisk variabel X med **tæthedsfunktion**

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

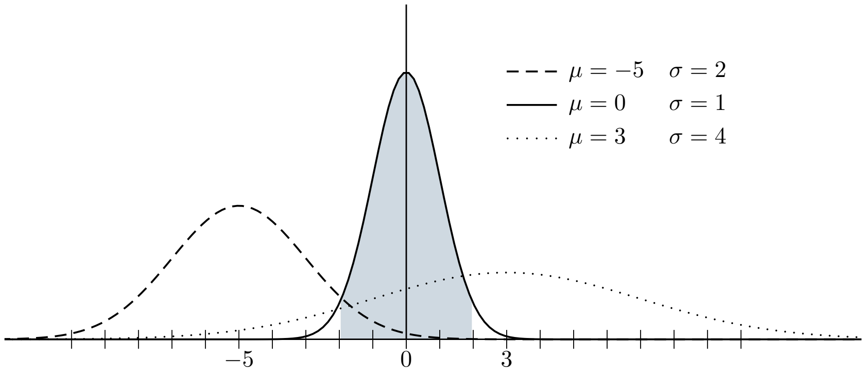
siges at være **normalfordelt** med middelværdi μ og varians σ^2, hvor μ og σ er reelle tal og $\sigma > 0$.

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. `matlab` (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

$$P(X \le x) = P\Big(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\Big) = P\Big(Z \le \frac{x-\mu}{\sigma}\Big) = \Phi\Big(\frac{x-\mu}{\sigma}\Big).$$

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 $P(X \le 5)$ i Matlab:

```
>> 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 $P(Z \le z) = 0{,}8$.
Løsning:

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

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


Opsummering: sandsynligheder og fraktiler for normalfordeling

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

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

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

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

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

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

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

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

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