

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

Svante Eriksen

Institut for Matematiske Fag

svante@math.aau.dk

Litteratur

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

Kursusform

Fem kursusgange, hver med vekslen mellem forelæsninger og opgaver.

Eksamen

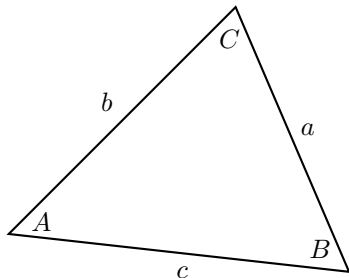
Mundtlig og fælles med *Det matematiske grundlag for kortprojektioner.*
Tager udgangspunkt i 4 opgaver - 2 fra hvert af de to kursusdele.

Formål og indhold

- 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.

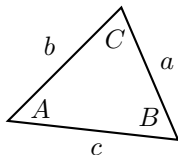
Formål og indhold



Arealet af ovenstående trekant kan bestemmes vha: $\text{areal} = \frac{1}{2}a \cdot b \cdot \sin C$.

Hvordan kvantificerer vi usikkerheden på målinger/skøn af a , b og C ?
Hvordan kvantificerer vi den resulterende usikkerhed på det beregnede areal ?

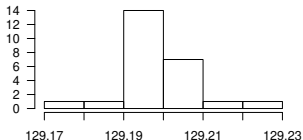
Formål



Antag vi har følgende 25 målinger af C :

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

Histogram - hvad er det? - over målingerne:



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

Kender vi usikkerheden på skønnene/estimerne af a , b og C giver **Fejlförplantningsloven** et skøn over usikkerheden på det beregnede areal.

Så langt når vi ikke på dette semester, men vi vender tilbage til problemstillingen på 5. semester.

Lige nu koncentrerer vi os om følgende:

- a og b antages kendte.
- Hvordan forplanter en fejl på C sig til arealberegningen?

Det skal vi lære at kende som **Den simple fejlförplantningslov**.

Enggaard A/S

Enggaard A/S vil gerne have:

- bedste bud på en given længde.
- Et interval af mulige værdier under hensyntagen til usikkerhed på måling.
- Krav: fejlen skal være mindre end f.eks. 5cm.

Vi lærer i dette kursus at håndtere ovennævnte...

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 $P(H) = 0,5$.

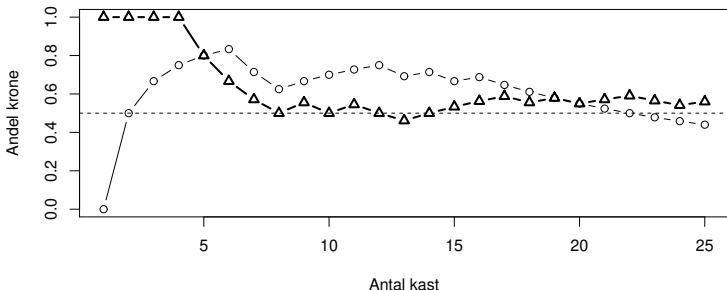
Eksempel - fortsat

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 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

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

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 (f.eks. måling af vinkel, længde, højdeforskel, etc), 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 !

Kontinuert og diskret stokastisk variabel

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

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.

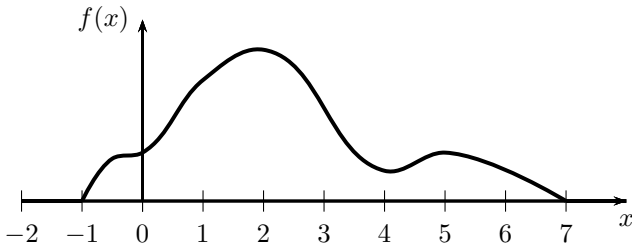
Tæthedsfunktion

Definition: Tæthedsfunktion

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

1. $f(x) \geq 0$ for alle $x \in \mathbb{R}$.
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 sandsynligheden for at X ligger mellem a og b ($a \leq b$) givet ved

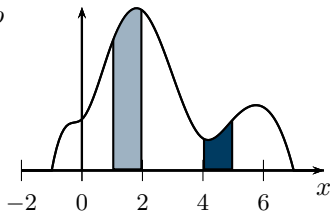
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

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

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

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

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



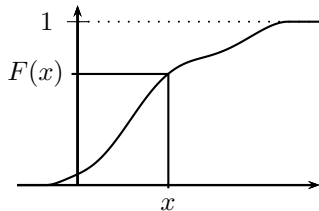
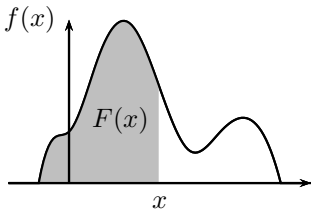
Fordelingsfunktionen

Definition: Fordelingsfunktion

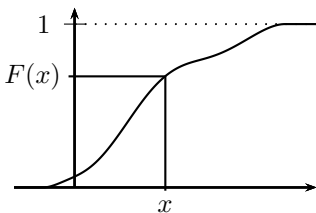
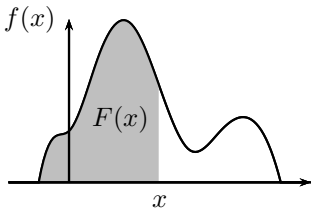
Lad X være en stokastisk variabel med tæthedsfunktion $f(x)$. Da er fordelingsfunktionen $F(x)$ givet ved

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

Eksempel på tæthedsfunktion og tilhørende fordelingsfunktion:



Fordelingsfunktionen: Egenskaber



Fordelingsfunktionen $F(x)$ har følgende egenskaber:

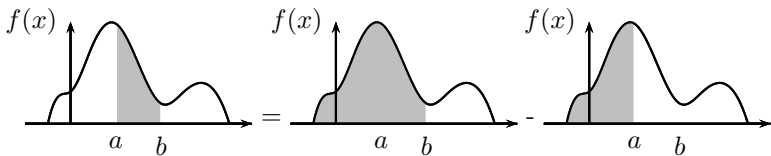
- $F(x)$ er en ikke-aftagende funktion
- $\lim_{x \rightarrow -\infty} F(x) = 0$
- $\lim_{x \rightarrow \infty} F(x) = 1$
- $F'(x) = f(x)$

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, \dots repræsentere gentagne målinger af samme størrelse under ens betingelser.

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

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

hvor x_1, x_2, \dots er udfald af X_1, X_2, \dots

Middelværdi - definition ud fra tæthed

Definition: Middelværdi

Lad X være en kontinuert stokastisk variabel med tæthedsfunktion f .

Middelværdien af X defineres som

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

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

Bemærk: $f(x)dx \approx P(x \leq X \leq x + dx)$.

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

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

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

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

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:

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

Eksempel: Telefonabonnement

Den månedlige samtale tid 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 samtale tid 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

Lad X være en kontinuert stokastisk variabel med tæthedsfunktion f og middelværdi μ .

Det er relevant at vide, hvor præcist X måler μ - som fex er den "sande" vinkel.

Fejlen på målingen er

$$\blacksquare \varepsilon = X - \mu$$

Denne er i gennemsnit lig med nul. Så når vi skal have et mål for afvigelsen fra μ kigger vi på **den kvadratiske fejl**

$$\blacksquare \varepsilon^2 = (X - \mu)^2$$

Den **forventede** værdi af den kvadratiske fejl kaldes variansen og betegnes σ^2 , dvs

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

Varians - frekventiel definition

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

Så svarer den fælles varians $\sigma^2 = \text{Var}X_1 = \text{Var}X_2 = \dots$ til grænseværdien af det empiriske gennemsnit af kvadratiske fejl:

$$\frac{1}{n}((x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2) \rightarrow \sigma^2 \quad \text{når } n \rightarrow \infty$$

hvor x_1, x_2, \dots er udfald af X_1, X_2, \dots

Formel definition - når $n = \infty$: Varians

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

$$\sigma^2 = \text{Var}(X) = \mathbb{E} [(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx < \infty,$$

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:

$$\sigma = \sqrt{\text{Var}(X)}$$

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

$$\text{Var}[Y] = \text{Var}[aX + b] = a^2 \text{Var}[X] = a^2 \sigma^2.$$

Bevis:

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

Bemærk at b ikke indgår.

Eksempel: Telefonabonnement

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

Hvad er variansen for den månedlige udgift?

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

Variansen er altså 2,96 kr².

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

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).