Noter i fejlteori

Kasper Klitgaard Berthelsen
Poul Winding
&
Jens Møller Pedersen

Diverse opdateringer ved Rasmus Waagepetersen.

Version 1.3

April 2016

Indhold

Forord

Denne note er en omarbejdning og udvidelse af en mere end 20 år gammel note "Noter i fejlteori" af Poul Winding og Jens Møller Pedersen. En særlig tak skal gå til folk der hjulet med tilblivelsen af denne note. Daniel Philip Holt for at have assisteret med konverteringen af den orginale note fra fotokopi til LaTeX. Malene Ravn for at have nærlæst version 1.1 af noten og påpeget fejl og mulige forbedringer.

Beviser afsluttes med en firkant $\square$.
Definitioner afsluttes med en trekant $\triangle$.
Eksempler afsluttes med en rombe $\lozenge$.

Kasper K. Berthelsen
Aalborg, 2014

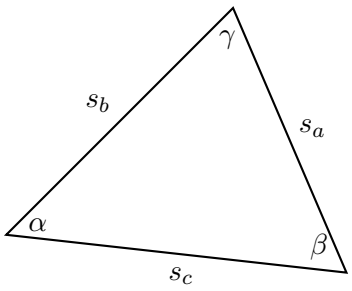
Rettelser af diverse trykfejl samt notation.
Rasmus Waagepetersen
Aalborg, 2016.

Kapitel 1

Motivation

Antag, at vi er interesserede i arealet af trekanten i figur 1.1. Hvis vi kender vinklen α og længderne s_b og s_c, kan vi bestemme arealet af trekanten vha. arealformlen

$$\text{areal} = \frac{1}{2} s_b s_c \sin(\alpha). \tag{1.1}$$

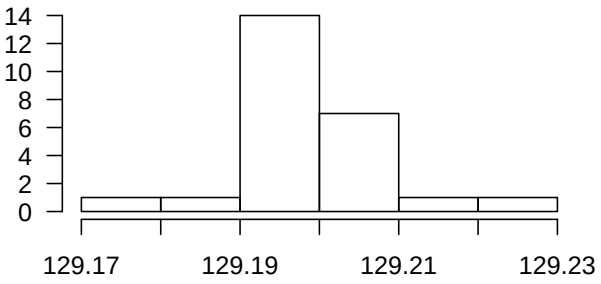


Figur 1.1: En trekant.

For at gøre situationen lidt simplere, antager vi, at vi kender længderne s_b og s_c. Det eneste, vi skal måle, er således vinklen α. Antag, at vi har følgende 25 målinger af α:

129.188, 129.203, 129.211, 129.177, 129.204, 129.205, 129.194, 129.195, 129.194, 129.191, 129.195, 129.19, 129.192, 129.201, 129.21, 129.199, 129.195, 129.191, 129.192, 129.224, 129.201, 129.195, 129.196, 129.205, 129.193.

Målingerne er også illustreret i histogrammet i figur 1.2. Gennemsnittet af de 25 målinger er 129.1976. Bemærk, hvordan målingerne er koncentreret omkring gennemsnittet. Hvad er på baggrund af målingerne et godt bud på den *sande vinkel*? Det viser sig, at gennemsnittet af målingerne er et godt bud. Som det tydeligt fremgår, så er målingen af α forbundet med en vis usikkerhed. Hvordan skal vi



Figur 1.2: Histogram for de 25 målinger af α.

opgøre denne usikkerhed? Kan vi sige noget om, hvor tæt gennemsnittet er på den sande vinkel? Da bestemmelsen af α på baggrund af målinger er forbundet med en vis usikkerhed, vil bestemmelsen af arealet af trekanten også være forbundet med en vis usikkerhed. Spørgsmålet er, hvordan usikkerheden for α påvirker usikkerheden af arealet?

Kapitel 2

Det matematiske fundament

2.1 Lidt sandsynlighedsregning

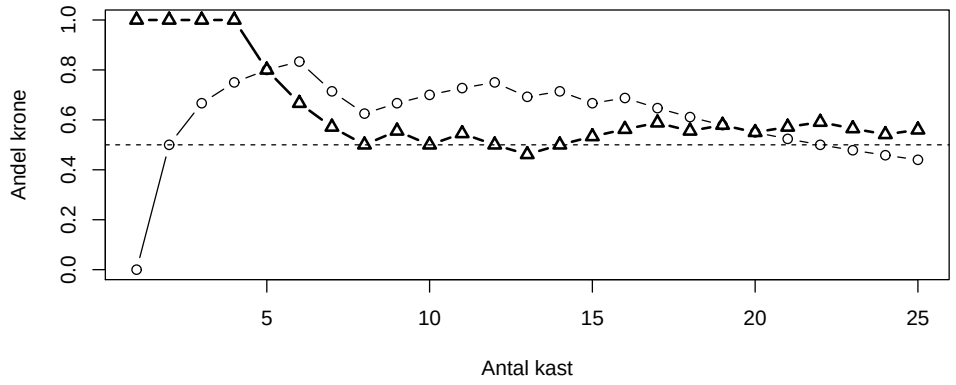
Når man foretager en måling af fx. en længde, er det forbundet med en vis usikkerhed. Et naturligt spørgsmål er, hvor stor sandsynligheden er, for at målingen ligger indenfor en vis fejlmargin fra den sande længde. For at kunne svare på dette må vi starte med at præcisere, hvad vi mener med sandsynlighed. Desuden skal det formuleres præcist, på hvilken måde målingerne er usikre.

Betragt et eksperiment, der kan ende i et eller flere *udfald*. Eksperimentet kunne være at måle en given længde, hvorved udfaldet er den faktisk målte længde. Et andet eksperiment kunne være at kaste en mønt, hvor de mulige udfald er plat og krone. Vi starter med det simple tilfælde, hvor eksperimentet har to udfald, som vi betegner som *succes* og *fiasko*. Eksperimentet kunne være at kaste med en mønt, hvor krone svarer til succes og plat svarer til fiasko. Antag, at vi gentager nøjagtigt det samme eksperiment igen og igen, og hvert eksperiment udføres uafhængigt af tidligere eksperimenter. Sandsynligheden for succes er da andelen af succeser i det lange løb. Sandsynligheden for succes betegnes $P(succes)$. Det ses umiddelbart, at $0 \leq P(succes) \leq 1$. Da hvert eksperiment enten er en succes eller en fiasko, følger det, at sandsynligheden for fiasko er $P(fiasko) = 1 - P(succes)$.

Figur 2.1 viser to eksempler på, hvordan andelen af krone løbende udvikler sig efterhånden som en fair mønt kastes flere og flere gange. Da mønten er fair, er andelen af krone i det lange løb 0,5, dvs. $P(krone) = 0,5$. Af figuren ses det, at den observerede andel i begge tilfælde nærmer sig 0,5.

Mere generelt kan vi antage, at et eksperiment kan resultere i en lang række udfald. Som nævnt kunne eksperimentet være at måle en længde. De mulige udfald er da alle positive reelle tal.

En *hændelse* er en mængde af udfald. Fx. kunne en hændelse være, at den målte længde er mellem 4,71 og 4,73 meter. Sandsynligheden for denne hændelse betegnes P(måling ligger mellem 4,71 og 4,73 meter). Sandsynligheden betegner andelen af målinger, der ligger mellem 4,71 og 4,73, hvis vi bliver ved med at måle længden igen og igen. Hver måling skal være foretaget under nøjagtig samme



Figur 2.1: Løbende andel af krone i to følger af uafhængige kast med en fair mønt.

betingelser og upåvirket af tidligere målinger.

Sætning 1 (Egenskaber for sandsynligheder)
Lad H være en vilkårlig hændelse, da gælder der følgende regler

1. $0 \leq P(H) \leq 1$
2. $P(ej\ H) = 1 - P(H)$.

Hændelse "ej H" betegnes den *komplementære* hændelsen til H, da hændelse "ej H" indtræffer, hvis H ikke indtræffer.

2.2 Stokastiske variable

I det følgende er udgangspunktet, at vi udfører et eksperiment, hvor udfaldet af eksperimentet kan konverteres til et reelt tal X. Som eksempel kunne eksperimentet være at måle vinklen α og lade X betegne målingen. Som indikeret af målingerne i Kapitel 1 er X tilfældig. Vi betegner derfor X en *stokastisk variabel*. Vi vil betragte en stokastisk variabel som en matematisk model for en måling, der er behæftet med en (tilfældig) fejl. I denne note vil vi kun betragte stokastiske variable, der kan tage alle værdier i et interval på den reelle akse. Fx. hvis X svarer til måling af en længde, da er $X \in [0, \infty)$. Hvis X er en vinkel målt i gon, da er $X \in [0, 400)$.

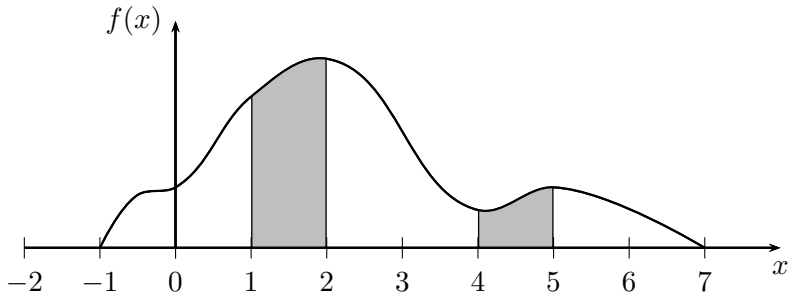
For at beskrive, på hvilken måde målingen X er tilfældig, har vi brug for at definere en såkaldt *tæthedsfunktion*.

Definition 1 (Tæthedsfunktion)
En tæthedsfunktion $f(x)$ er en reel funktion, der opfylder

1. $f(x) \geq 0$ for alle $x \in R$.

2. $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$. △

En tæthedsfunktion er altså en ikke-negativ funktion, hvor det totale areal under funktionen er 1. Figur 2.2 viser et eksempel på en tæthedsfunktion



Figur 2.2: Eksempel på tæthedsfunktion.

Tæthedsfunktioner bruges til at beskrive fordelingen af en stokastisk variabel:

Definition 2 (Tæthedsfunktion og stokastisk variabel)
En stokastisk variabel X har tæthedsfunktion f, hvis det for alle reelle tal a og b, hvor $a \leq b$, gælder, at sandsynligheden for, at X ligger i intervallet fra a til b er givet ved

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x. \tag{2.1}$$

Da $f(x)$ er ikke-negativ, er sandsynligheden for at X ligger mellem a og b givet ved arealet fra a til b under grafen $f(x)$ (og over første-aksen).

Arealet af det skraverede område til venstre i figur 2.2 svarer til sandsynligheden for, at X ligger i intervallet mellem 1 og 2. Ligeledes svarer det skraverede område til højre i figur 2.2 til sandsynligheden for, at X ligger i intervallet mellem 4 og 5. Med udgangspunkt i figur 2.2 er det klart, at $P(1 \leq X \leq 2)$ er større end $P(4 \leq X \leq 5)$, dvs. der er større sandsynlighed for, at X ligger mellem 1 og 2 end at X ligger mellem 4 og 5.

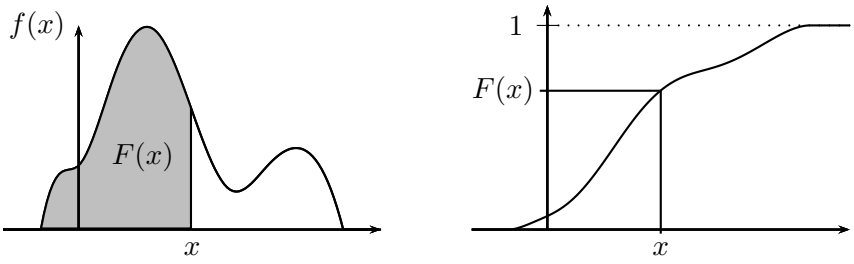
Sandsynligheder for en stokastisk variabel kan ækvivalent angives vha. *fordelingsfunktionen*.

Definition 3 (Fordelingsfunktionen)
Hvis X er en stokastisk variabel med tæthedsfunktion f, så er den tilsvarende fordelingsfunktion givet ved

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

Der gælder da omvendt, at $F'(x) = f(x)$, dvs. tæthedsfunktionen er den afledte af fordelingsfunktionen.

Figur 2.3 viser et eksempel på en tæthedsfunktion og den tilsvarende fordelingsfunktion. Fordelingsfunktionen $F(x)$ svarer til det grå område i det venstre plot i figur 2.3. Da fordelingsfunktionen er et integral af en ikke-negativ funktion, følger det, at fordelingsfunktionen $F(x)$ er en ikke-aftagende funktion. Desuden gælder der at $\lim_{x\to\infty} F(x) = 1$ og $\lim_{x\to-\infty} F(x) = 0$.



Figur 2.3: Til venstre: eksempel på tæthedsfunktion $f(x)$. Til højre: tilsvarende fordelingsfunktion $F(x)$.

Sandsynligheden (2.1) kan nu skrives som

$$P(a \leq X \leq b) = F(b) - F(a). \tag{2.2}$$

Fordelen ved denne formel fremfor (2.1) er at $F(x)$ ofte er tilgængelig enten i tabelform eller vha. software.

At (2.2) er korrekt ses af følgende udregning

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x)\mathrm{d}x \\ &= \int_{-\infty}^b f(x)\mathrm{d}x - \int_{-\infty}^a f(x)\mathrm{d}x \\ &= F(b) - F(a). \end{aligned}$$

I de fleste praktiske problemstillinger består en måleopgave i mere end en måling. Vi vil i det efterfølgende ofte antage, at den tilfældige fejl i en måling ingen indflydelse har på den tilfældige fejl i en anden måling. Mere præcist antager vi at to målinger er uafhængige:

Definition 4 (Uafhængighed)
To stokastiske variable X og Y kaldes *uafhængige*, hvis og kun hvis $P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$ for alle reelle tal a og b. △

Øvelse 1
Betragt følgende funktion

$$f(x) = \begin{cases} x & \text{når } 0 \leq x \leq 1 \\ 2 - x & \text{når } 1 < x \leq 2 \\ 0 & \text{ellers.} \end{cases}$$

1. *Skitser funktionen $f(x)$.*
2. *Opfylder $f(x)$ kravene til en tæthedsfunktion?*

2.3 Middelværdi og varians

To vigtige karakteristika for alle stokastiske variable er deres middelværdi og varians. Hvis man tænker på den stokastiske variabel X som en model for en måling, er middelværdien for den stokastisk variabel X gennemsnittet af målingerne i det lang løb. Variansen er et udtryk for, hvor meget den stokastiske variabel varierer omkring middelværdien. Hvis X repræsenterer en måling, kan variansen ses som et mål for kvaliteten af målingen. Jo mindre varians jo mindre variation fra måling til måling, hvilket kan tages som et udtryk for en højere kvalitet af målingen.

Definition 5 (Middelværdi)
Middelværdien for en stokastisk variabel X betegnes $\mathbb{E}[X]$ og er givet ved

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x. \tag{2.3}$$

Nogle gange omtales middelværdien for en stokastisk variabel også *forventningen* eller den forventede værdi. På engelsk bliver det til expectation, hvilket er forklaringen på, at bogstavet $\mathbb{E}$ benyttes til at betegne middelværdi for en stokastisk variabel.

I mange tilfælde er vi ikke interesseret i middelværdien for X, men derimod for middelværdien af en anden størrelse, som er en funktion af X:

Definition 6
Hvis $h(x)$ er en reel funktion, er middelværdien af $h(X)$ givet ved

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x. \tag{2.4}$$

Udregningen af $\mathbb{E}[h(X)]$ kan i almindelighed være vanskelig. En undtagelse er de tilfælde, hvor $h(x)$ er en lineær funktion.

Sætning 2 (Middelværdien for en lineær transformation)
Antag, at X er en stokastisk variabel med middelværdi $\mathbb{E}[X] = \mu$. Da er middelværdien for transformationen $aX + b$ givet ved

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

Bevis Antag at X er en stokastisk variabel med tæthedsfunktion f og $h(x) = ax + b$. Da følger $\mathbb{E}[aX + b]$ af (2.4):

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

hvor vi har benyttet definition 6 og egenskab 2 i definition 1. □

Hvis vi kender $\mathbb{E}[X]$, er det med andre ord en simpel opgave at finde $\mathbb{E}[a + bX]$.

Det er nemt at udvide sætning 2 til middelværdien for en linearkombination af flere stokastiske variable:

Sætning 3
Lad $X_1, X_2, \ldots, X_n$ være n stokastiske variable med middelværdierne $\mu_1, \ldots, \mu_n$, dvs. $\mathbb{E}[X_i] = \mu_i$. For alle reelle tal $a_0, a_1, a_2, \ldots, a_n$ gælder, at middelværdien af linearkombinationen $a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$ er givet ved

$$\mathbb{E}[a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_0 + a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n.$$

Bemærk, at der ikke er nogen antagelser om, at $X_1, \ldots, X_n$ er indbyrdes uafhængige stokastiske variable.

Variansen for en stokastisk variabel X er et udtryk for, hvor meget en stokastisk variabel varierer omkring middelværdien. Mere præcist definerer vi variansen som:

Definition 7 (Variansen)
Variansen for en stokastisk variabel X med middelværdi μ er defineret som

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[(X - \mu)^2], \tag{2.5}$$

dvs. middelværdien af den kvadrerede afstand mellem X og middelværdien μ.△

Bemærk, at definition af varians svarer til (2.4), hvor $h(x) = (x-\mu)^2$. Hvis X har tæthedsfunktion $f(X)$, kan variansen for X udregnes vha.

$$\mathbb{V}\text{ar}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

I forbindelse med praktiske udregninger af variansen er det ofte bekvemt at benytte følgende omskrivning af (2.5):

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mu^2. \tag{2.6}$$

Denne sammenhæng kan nemt udledes:

$$\begin{aligned}\mathbb{V}\text{ar}[X] &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[X^2 + \mu^2 - 2\mu X] \\ &= \mathbb{E}[X^2] + \mu^2 - 2\mu\mathbb{E}[X] \\ &= \mathbb{E}[X^2] - \mu^2.\end{aligned}$$

Definition 8 (Standardafvigelsen)
Standardafvigelsen σ er kvadratroden af variansen. △

Dvs. hvis den stokastiske variabel X har varians σ^2, er standardafvigelsen for X givet ved $\sigma = \sqrt{\sigma^2}$.

Det er generelt svært at finde variansen for $h(X)$ på nær, når $h(x) = ax + b$:

Sætning 4
Antag, at X er en stokastisk variabel med middelværdi $\mathbb{E}[X] = \mu$ og varians $\mathbb{V}\mathrm{ar}[X] = \sigma^2$. Da er variansen af transformationen $aX + b$ givet ved

$$\mathbb{V}\mathrm{ar}[aX + b] = a^2\mathbb{V}\mathrm{ar}[X]. \tag{2.7}$$

Bevis Husk at $\mathbb{V}\mathrm{ar}[h(X)] = \mathbb{E}[(h(X) - E[h(X)])^2]$. Hvis vi antager, at $h(X) = aX + b$ og bemærker, at $\mathbb{E}[aX + b] = a\mu + b$ så har vi

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

□

Bemærk, at konstanten b ikke optræder i resultatet (2.7). Intuitionen er, at b blot bidrager med en forskydning af X og dermed ikke påvirker variationen.

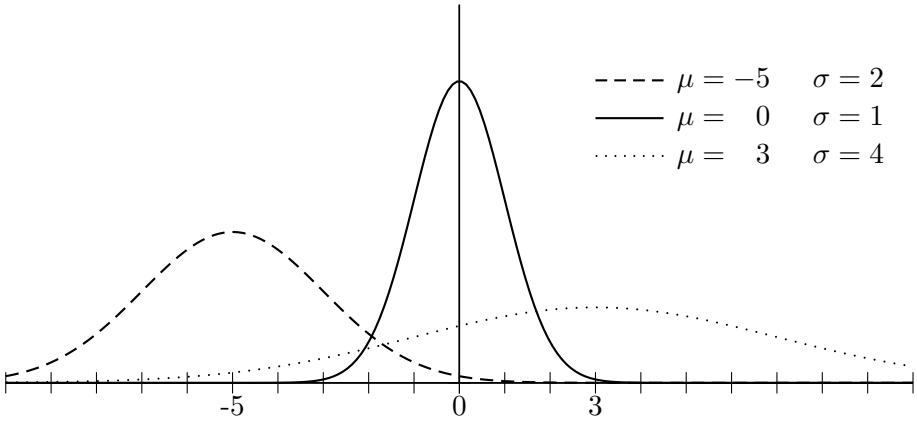
Sætning 4 kan udvides til at gælde en generel linearkombination af flere stokastiske variable:

Sætning 5
Lad $X_1, X_2, \ldots, X_n$ være n uafhængige *stokastiske variable med varianser $\sigma_1^2, \ldots, \sigma_n^2$, dvs. $\mathbb{V}\mathrm{ar}[X_i] = \sigma_i^2$. For alle reelle tal $a_0, a_2, \ldots, a_n$ gælder, at variansen af linearkombinationen $a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$ er givet ved*

$$\mathbb{V}\mathrm{ar}[a_0 + a_1X_1 + a_2X_2 + \cdots + a_kX_k] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_k^2\sigma_k^2. \tag{2.8}$$

Bemærk: en forudsætning for sætning 5 er, at de n stokastiske variable er indbyrdes uafhængige. Denne begrænsning vil vi råde bod på senere.

Eksempel 1
Fridjof er verdens kedeligste frugthandler! Han sælger *kun* æbler og pærer fra sin lille bod på torvet. For hvert æble tjener han 1,27 kr og for hver pære tjener han 0,87. Desuden koster det ham 119 kr. om dagen i faste udgifter at drive boden på torvet. Vi ved desuden, at det forventede antal solgte æbler og pærer er hhv. 97,3 og 63,4. Hvad er det forventede daglige overskud for Fridjof?
Løsning: Lad X og Y være stokastiske variable, der svarer til det solgte antal hhv. æbler og pærer. Dvs. $\mathbb{E}[X] = 97,3$ og $\mathbb{E}[Y] = 63,4$. Det daglige overskud



Figur 2.4: Tre eksempler på normalfordelinger.

betegner vi $S = 1,27X + 0,87Y - 119$. Det forventede overskud er derfor $\mathbb{E}[S]$, som vi udregner:

$$\begin{aligned}\mathbb{E}[S] &= \mathbb{E}[1,27X + 0,87Y - 119] \\ &= 1,27\mathbb{E}[X] + 0,87\mathbb{E}[Y] - 119 \\ &= 1,27 \cdot 97,3 + 0,87 \cdot 63,4 - 119 \\ &= 59,729.\end{aligned}$$

Dvs. det forventede daglige overskud er knap 60kr. ◊

2.4 Normalfordelingen

Den måske vigtigste fordeling af alle fordelinger er *normalfordelingen*.

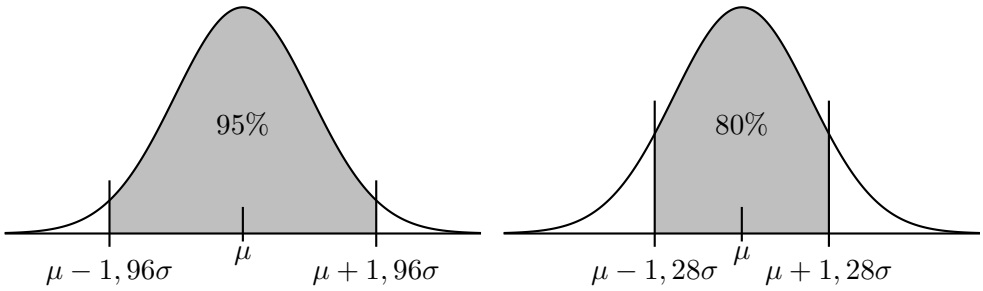
Definition 9 (Normalfordelingen)
En stokastisk variabel med tæthedsfunktion

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

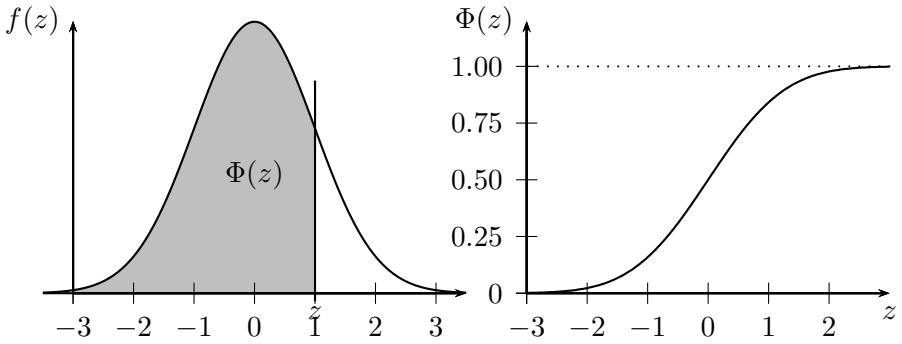
kaldes normalfordelt med middelværdi μ og varians σ^2. Som bekvem notation benyttes $X \sim \mathcal{N}(\mu, \sigma^2)$, der skal læses som 'den stokastiske variabel X følger en normalfordeling med middelværdi μ og varians σ^2'. △

Normalfordelingen er en god (tilnærmet) beskrivelse af mange i praksis forekommende tilfældige fænomener. Figur 2.4 viser tre eksempler på normalfordelinger. Bemærk, at tæthedsfunktionen er symmetrisk omkring middelværdien μ.

Antag $X \sim \mathcal{N}(\mu, \sigma^2)$, dvs. X er en normalfordelt stokastisk variabel med middelværdi μ og varians σ^2. Uanset hvilken middelværdi μ og varians σ^2 man vælger,



Figur 2.5: Venstre plot: Tæthedsfunktion for normalfordelt stokastisk variabel med forventning μ og varians σ^2. Den normalfordelte stokastiske variabel ligger med 95% sandsynlighed i intervallet $\mu \pm 1,96\sigma$ svarende til arealet af det grå område i figuren. Højre plot: Som venstre plot, men her illustrerer det grå område, at der er 80% sandsynlighed for, at den normalfordelte stokastiske variabel ligger i intervallet $\mu \pm 1,28\sigma$.



Figur 2.6: Venstre plot: Tæthedsfunktionen for en standard normalfordelt stokastisk variabel. Det grå område illustrerer fordelingsfunktionen $\Phi(z) = P(Z \leq z)$. Højre plot: Fordelingsfunktionen for en standard normalfordelt stokastisk variabel.

gælder der altid, at sandsynligheden for at X ligger højst 1,96 standardafvigelser fra middelværdien er 95%. Dette er illustreret på den venstre graf i figur 2.5. Tilsvarende er der 80% sandsynlighed for at X ligger højst 1,28 standardafvigelser fra middelværdien. Dette er illustreret i den højre graf i figur 2.5. Disse egenskaber vender vi tilbage til i sætning 8 og i eksempel 11.

Normalfordelingen med middelværdi 0 og varians 1 betegnes *standard normalfordelingen*. I denne note vil vi typisk betegne en stokastisk variabel, der følger en standard normalfordeling ved Z, dvs. $Z \sim \mathcal{N}(0;1)$. Desuden betegnes fordelingsfunktionen for standard normalfordelingen i denne note ved $\Phi(z)$. Hvis $Z \sim \mathcal{N}(\mu;\sigma^2)$, har vi derfor $P(Z \leq z) = \Phi(x)$. Tæthedsfunktionen for standard normalfordelingen og dens fordelingsfunktion Φ er illustreret på figur 2.6. Fordelingsfunktionen Φ er tabellagt i tabel 6.1 på side 72. Desuden har de fleste typer software (fx. Matlab), der kan benyttes til statistik, funktioner indbygget, der kan

udregne fordelingsfunktionen for en vilkårlig normalfordeling (og dermed også for standard normalfordelingen).

Eksempel 2
Antag, at Z er standard normalfordelt, dvs. $Z \sim \mathcal{N}(0,1)$. Vi ønsker nu at finde følgende sandsynligheder $P(Z \leq 1,17)$ og $P(Z \leq -1,82)$.
Løsning: Den første sandsynlighed,$P(Z \leq 1,17)$, findes umiddelbart vha. normalfordelingstabellen: $P(Z \leq 1,17) = \Phi(1,17) = 0,8790$.

Den næste sandsynlighed, $P(Z \leq -1,82)$, kan ikke umiddelbart slås op, da $\Phi(z)$ kun er tabellagt for positive værdier af z. Da standard normalfordelingen er symmetrisk omkring nul, gælder $P(Z \leq -1,82) = P(Z \geq 1,82)$. Ifølge egenskab 2 ved sandsynligheder har vi $P(Z \geq 1,82) = 1 - P(Z \leq 1,82) = 1 - \Phi(1,82) = 1 - 0,9656 = 0,0344$. Dvs. $P(Z \leq -1,82) = 0,0344$. $\Diamond$

I eksempel 2 benytter vi sammenhængen $\Phi(-1,82) = 1 - \Phi(1,82)$. Denne sammenhæng gælder helt generelt:

Sætning 6
Antag, at $Z \sim \mathcal{N}(0,1)$, dvs. Z er standard normalfordelt. Da gælder, at $\Phi(Z \leq -z) = 1 - P(Z \leq z)$, hvilket kan skrives som $\Phi(-z) = 1 - \Phi(z)$.

Bevis Da standard normalfordelingen er symmetrisk omkring nul, gælder der $P(Z \leq z) = P(Z \geq -z)$.

$$\begin{aligned} P(Z \leq -z) &= 1 - P(Z \geq -z) \\ &= 1 - P(Z \leq z) \end{aligned}$$

Den sidste ligning kan også skrives som $\Phi(-z) = 1 - \Phi(z)$. $\square$

Ud over at være en god beskrivelse af mange virkelige problemstillinger, så har normalfordelingen mange nyttige egenskaber. Én egenskab er, at en lineær transformation af en normalfordelt stokastisk variabel også er normalfordelt:

Sætning 7
Antag, at a og b er reelle konstanter, $X \sim \mathcal{N}(\mu, \sigma^2)$ og $Y = aX + b$. Da gælder $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

At Y har middelværdi $a\mu + b$ og varians $a^2\sigma^2$ er ikke overraskende, da det følger af sætningerne 2 og 4. Det interessante er, at en lineær transformation af X også er normalfordelt.

Eksempel 3 (Standardisering)
Antag, at $X \sim \mathcal{N}(\mu; \sigma^2)$ og definer

$$Z = \frac{X - \mu}{\sigma}. \tag{2.9}$$

Hvilken fordeling følger Z?
Løsning: Bemærk at Z kan skrives som $Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$. Sætter vi $a = \frac{1}{\sigma}$ og $b = -\frac{\mu}{\sigma}$ følger det af sætning 7, at Z er normalfordelt med middelværdi 0 og varians 1, dvs. Z følger en standard normalfordeling. Vi siger, at vi har *standardiseret* den oprindelige stokastiske variabel X. ◊

Eksempel 2 omhandlede sandsynligheder for en standard normalfordelt stokastisk variabel. Alle udregninger involverede tabelopslag af $\Phi(z)$. Hvis en stokastisk variabel X er normalfordelt, men ikke standard normalfordelt, kan vi ikke direkte finde sandsynligheder vha. $\Phi(z)$ — men vha. en standardisering er det muligt. Hvis X er normalfordelt med middelværdi μ og varians σ^2, gælder der, at

$$\begin{aligned} P(X \leq x) &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right), \end{aligned} \tag{2.10}$$

hvor $Z \sim \mathcal{N}(0,1)$ og den første lighed følger af sædvanlige regneregler for uligheder idet $\sigma > 0$. En anden konsekvens af (2.10) er, at hvis $X \sim \mathcal{N}(\mu, \sigma^2)$, så er Xs fordelingsfunktion

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{2.11}$$

Eksempel 4
Antag $X \sim \mathcal{N}(3, 16)$, dvs. X er en normalfordelt stokastisk variabel med middelværdi 3 og varians 16. Find følgende sandsynligheder $P(X \leq 4,6)$, $P(X \leq 2,2)$ og $P(X \geq 2,8)$.
Løsning: Det følger af standardiseringen (2.9) i eksempel 3, at $Z = (X - 3)/\sqrt{16} \sim \mathcal{N}(0,1)$.

$$\begin{aligned} P(X \leq 4,6) &= P\left(\frac{X-3}{4} \leq \frac{4,6-3}{4}\right) \\ &= P(Z \leq 0,4) = \Phi(0,4) = 0,6554. \end{aligned}$$

Dvs. sandsynligheden for at X er højst 4,6 er omtrent 65,5%.

$$\begin{aligned} P(X \leq 2,2) &= P\left(\frac{X-3}{4} \leq \frac{2,2-3}{4}\right) = P(Z \leq -0,2) \\ &= \Phi(-0,2) = 1 - \Phi(0,2) = 1 - 0,5793 = 0,4207. \end{aligned}$$

Dvs. sandsynligheden for at X er højst 2,2 er omtrent 42%.

$$\begin{aligned} P(X \geq 2,8) &= P\left(\frac{X-3}{4} \geq \frac{2,8-3}{4}\right) = P(Z \geq -0,05) = 1 - P(Z \leq -0,05) \\ &= 1 - \Phi(-0,05) = 1 - (1 - \Phi(0,05)) = 0,5199. \end{aligned}$$

Dvs. sandsynligheden for at X er mindst 2,8 er omtrent 52%. ◊

I eksemplerne ovenfor udregner vi sandsynligheden for, at en normalfordelt stokastisk variabel ligger over (eller under) en bestemt værdi. I nogle tilfælde kan vi være interesseret i at finde en værdi, som den normalfordelte stokastiske variable ligger under eller over med en given sandsynlighed. Til dette formål har vi brug for den *inverse fordelingsfunktion* for en standard normalfordeling, betegnet Φ^{-1}. At Φ^{-1} er invers til Φ betyder at $\Phi^{-1}(\Phi(z)) = z$ og $\Phi(\Phi^{-1}(p)) = p$. Vi kan finde værdier af $\Phi^{-1}(p)$ vha. tabel 6.1 ved først at finde den sandsynlighed der er nærmest p og derefter vælge den tilhørende z værdi.

Eksempel 5
Antag, at Z er standard normalfordelt, dvs. $Z \sim \mathcal{N}(0,1)$. Vi ønsker nu at finde et tal z, så sandynligheden for at Z er mindre end eller lig dette tal er 72%, dvs. vi ønsker at finde z, så $P(Z \leq z) = 0,72$.
Løsning: Vi benytter regnereglen $\Phi^{-1}(\Phi(z)) = z$:

$$\begin{aligned} P(Z \leq z) &= 0,72 && \Leftrightarrow \\ \Phi(z) &= 0,72 && \Leftrightarrow \\ z &= \Phi^{-1}(0,72) \end{aligned}$$

Vi mangler kun at bestemme $\Phi^{-1}(0,72)$. Ifølge tabel 6.1 har vi $\Phi(0,58) = 0,7190$ og $\Phi(0,59) = 0,7224$. Dvs. $\Phi^{-1}(0,72)$ er et sted mellem 0,58 og 0,59. Da 0,7190 er nærmere 0,72 end 0,7224 vælger vi 0,58. Med andre ord $z = \Phi^{-1}(0,72) \approx 0,58$. ◊

Ved hjælp af tabel 6.1 er det ikke umiddelbart muligt at bestemme $\Phi^{-1}(p)$ når p er mindre end 0,5. Fra sætning 6 har vi $\Phi(z) = 1 - \Phi(-z)$. Benytter vi denne regneregel, har vi følgende omskrivning:

$$\begin{aligned} \Phi(z) &= p && \Leftrightarrow \\ 1 - \Phi(-z) &= && \Leftrightarrow \\ \Phi(-z) &= 1 - p && \Leftrightarrow \\ z &= -\Phi^{-1}(1-p) \end{aligned}$$

Hvis p er mindre ned 0,5, så er $1-p$ større end 0,5 og vi kan dermed bruge tabel 6.1 igen.

Eksempel 6
Antag, at Z er standard normalfordelt, dvs. $Z \sim \mathcal{N}(0,1)$. Vi ønsker nu at finde et tal z, så sandsynligheden for at Z er mindre end eller lig dette tal er 17%, dvs. vi ønsker at finde z, så $P(Z \leq z) = 0,17$.
Løsning: Vi benytter regnereglen $\Phi^{-1}(\Phi(z)) = z$:

$$\begin{aligned} P(Z \leq z) &= 0,17 && \Leftrightarrow \\ \Phi(z) &= 0,17 && \Leftrightarrow \\ z &= \Phi^{-1}(0,17) \end{aligned}$$

Vi kan ikke umiddelbart tilnærme $\Phi^{-1}(0,17)$ vha. tabel 6.1. I stedet bemærker vi, at $\Phi(z) = 1 - \Phi(-z)$. Omskrivningen bliver nu

$$z = -\Phi^{-1}(1 - 0,17) = -\Phi^{-1}(0,83)$$

Vi mangler kun at bestemme $\Phi^{-1}(0,83)$. Ifølge tabel 6.1 har vi $\Phi(0,95) = 0,8289$ og $\Phi(0,96) = 0,8315$. Dvs. $\Phi^{-1}(0,83)$ er et sted mellem 0,95 og 0,96. Da 0,8289 er nærmest 0,83, vælger vi 0,95. Dermed opnår vi $z = -\Phi^{-1}(0,83) \approx -0,95$. ◊

Vha. standardisering kan den inverse fordelingsfunktion også benyttes for andet end standard normalfordelte stokastiske variable:

Eksempel 7

Antag $X \sim \mathcal{N}(3,16)$, dvs. X er en normalfordelt stokastisk variabel med middelværdi 3 og varians 16. Find x så $P(X \leq x) = 0,87$.

Løsning: Det følger af standardiseringen (2.9) i eksempel 3, at $Z = (X - 3)/\sqrt{16} \sim \mathcal{N}(0,1)$.

$$\begin{aligned} P(X \leq x) &= 0,87 \\ P\left(\frac{X-3}{4} \leq \frac{x-3}{4}\right) &= 0,87 \\ \Phi(\frac{x-3}{4}) &= 0,87 \\ \frac{x-3}{4} &= \Phi^{-1}(0,87) \\ x &= 3 + 4\Phi^{-1}(0,87). \end{aligned} \tag{2.12}$$

Vi mangler kun at finde $\Phi^{-1}(0,87)$. Ifølge tabel 6.1 gælder der, at $\Phi(1,12) = 0,8686$ og $\Phi(1,13) = 0,8708$, dvs. $\Phi^{-1}(0,87)$ er et sted mellem 1,12 og 1,13. Vi vælger 1,13 da 0,8708 ligger nærmest 0,87. Indsættes 1,13 i (2.12) fås

$$x = 3 + 4 \cdot 1,13 = 7,52.$$

Dvs. der er (omtrent) 87% sandsynlighed for at X er højst 7,52. ◊

For en generel stokastisk variabel X med fordelingsfunktion F giver formel (2.2) at $P(a \leq X \leq b) = F(b) - F(a)$ for $a \leq b$. På tilsvarende vis kan fordelingsfunktionen Φ anvendes til at bestemme sandsynligheden for at en standard normalfordelt stokastisk variabel ligger i et givet interval.

Eksempel 8

Antag, at Z er standard normalfordelt, dvs. $Z \sim \mathcal{N}(0,1)$. Bestem sandsynligheden for at Z falder i intervallet fra -1 til 2, dvs. vi ønsker at bestemme $P(-1 \leq Z \leq 2)$.

Løsning: Da Φ er fordelingsfunktionen for Z følger det af (2.2):

$$\begin{aligned} P(-1 \leq Z \leq 2) &= \Phi(2) - \Phi(-1) = \Phi(2) - (1 - \Phi(1)) \\ &= 0,9772 - 1 + 0,8413 = 0,8185. \end{aligned}$$

Dvs. sandsynligheden for at en standard normalfordelt stokastisk variabel ligger i intervallet fra -1 til 2 er omtrent 81,9%. ◊

Hvis X er normalfordelt, men ikke standard normalfordelt gælder formel (2.2) selvfølgelig stadig. Vha. en standardisering er det muligt at bruge Φ:

Eksempel 9
Antag $X \sim \mathcal{N}(3, 16)$, dvs. X er en normalfordelt stokastisk variabel med middelværdi 3 og varians 16. Find sandsynligheden for at X ligger i intervallet fra 2,2 til 4,2, dvs. sandsynligheden $P(2,2 \leq X \leq 4,2)$.
Løsning: Det følger af standardiseringen (2.9) i eksempel 3, at $Z = (X - 3)/\sqrt{16} \sim \mathcal{N}(0,1)$.

$$\begin{aligned} P(2,2 \leq X \leq 4,2) &= P(X \leq 4,2) - P(X \leq 2,2) \\ &= P\left(\frac{X-3}{3} \leq \frac{4,2-3}{4}\right) - P\left(\frac{X-3}{3} \leq \frac{2,2-3}{4}\right) \\ &= P(Z \leq 0,3) - P(Z \leq -0,2) \\ &= \Phi(0,3) - \Phi(-0,2) = \Phi(0,3) - (1 - \Phi(0,2)) \\ &= 0,6179 - 1 + 0,5793 = 0,1972 \end{aligned}$$

Dvs. sandsynligheden for at X ligger i intervallet fra 2,2 til 4,2 er omtrent 19,7%. ◊

Sandsynligheden for at en normalfordelt stokastisk variabel ligger i et interval, der er symmetrisk omkring middelværdien, har særlig interesse. Længden af sådan et interval måles i standardafvigelser.

Sætning 8
Antag, at $X \sim \mathcal{N}(\mu, \sigma^2)$. Da gælder der

$$P(\mu - z\sigma \leq X \leq \mu + z\sigma) = 2\Phi(z) - 1.$$

Dvs. sandsynligheden for at X ligger højst z standardafvigelser fra middelværdien μ er givet ved $2\Phi(z) - 1$.

Bevis Resultatet følger af følgende omskrivninger, hvor vi benytter standardisering og fordelingsfunktionen for en standard normalfordelt stokastisk variabel:

$$\begin{aligned} P(\mu - z\sigma \leq X \leq \mu + z\sigma) &= P\left(-z \leq \frac{X-\mu}{\sigma} \leq z\right) \\ &= P(-z \leq Z \leq z) \\ &= P(Z \leq z) - P(Z \leq -z) \\ &= \Phi(z) - \Phi(-z) \\ &= \Phi(z) - (1 - \Phi(z)) \\ &= 2\Phi(z) - 1 \end{aligned}$$

Hermed er resultatet vist. □

Bemærk, at sandsynligheden for at en normalfordelt stokastisk variabel ligger højst z standardafvigelser fra middelværdien *ikke* afhænger af hverken middelværdien μ eller variansen σ^2.

Eksempel 10
Antag $X \sim \mathcal{N}(\mu, \sigma^2)$, dvs. X er normalfordelt med middelværdi μ og varians σ^2. Hvad er sandsynligheden for at X ligger højst 1,5 standardafvigelser fra middelværdien? Med andre ord, bestem sandsynligheden $P(\mu - 1,5\sigma \leq X \leq \mu + 1,5\sigma)$.
Løsning: Jf. sætning 8, så er sandsynligheden $2\Phi(1,5) - 1$. Fra tabel 6.1 har vi, at $\Phi(1,5) = 0,9332$, hvilket giver $2\Phi(1,5) - 1 = 2 \cdot 0,9332 - 1 = 0,8664$. Dvs. sandsynligheden for at en normalfordelt stokastisk variabel ligger højst 1,5 standardafvigelser fra middelværdien er omtrent 86,6%. ◊

I eksempel 10 ovenfor fandt vi sandsynligheden for, at en normalfordelt stokastisk variabel lå højst et givet antal standardafvigelser fra middelværdien. Næste eksempel omhandler den modsatte situation. Antag, at sandsynligheden for at ligge i intervallet er givet. Hvor mange standardafvigelser svarer det så til?

Eksempel 11
Antag $X \sim \mathcal{N}(\mu, \sigma^2)$, dvs. X er normalfordelt med middelværdi μ og varians σ^2. Find et z, så sandsynligheden for at X afviger højst z standardafvigelser fra middelværdien er 95%?
Løsning: Fra sætning 8 har vi, at sandsynligheden for, at en normalfordelt stokastisk variabel afviger med højst z standardafvigelser fra middelværdien er $2\Phi(z) - 1$. Spørgsmålet kan derfor formuleres som: find et tal z, så $2\Phi(z) - 1 = 0,95$. Da $\Phi^{-1}(\Phi(z)) = z$ har vi

$$\begin{aligned} 2\Phi(z) - 1 &= 0,95 && \Leftrightarrow \\ \Phi(z) &= \frac{0,95 + 1}{2} && \Leftrightarrow \\ z &= \Phi^{-1}(0,975) \end{aligned}$$

I tabel 6.1 ses, at $\Phi(1,96) = 0,9750$, så løsningen er $z = 1,96$. Dvs. for *alle* normalfordelte stokastiske variable gælder, at de med 95% sandsynlighed ikke afviger fra middelværdien med mere end 1,96 standardafvigelser. Dette resultat stemmer også fint overens med den venstre graf i figur 2.5. ◊

Fremgangsmåden i eksempel 11 kan nemt generaliseres. Traditionelt angives sandsynligheden for at lande i det symmetriske interval som $(1 - \alpha)100\%$. Antag derfor, at vi søger et z, så sandsynligheden for at X afviger højst z standardafvigelser fra middelværdien er $(1 - \alpha)$. Løsningen er

$$z = \Phi^{-1}(1 - \alpha/2). \tag{2.13}$$

Eksempel 12
Antag $X \sim \mathcal{N}(\mu, \sigma^2)$, dvs. X er normalfordelt med middelværdi μ og varians σ^2. Find et z, så sandsynligheden for at X afviger højst z standardafvigelser fra middelværdien er 80%?
Løsning: I dette tilfælde er $\alpha = 0,2$. Svaret er derfor, jf. formel (2.13), $z = \Phi^{-1}(1-0,2/2) = \Phi^{-1}(0,9)$. I tabel 6.1 finder vi $\Phi(1,28) = 0,8997$ og $\Phi(1,29) = 0,9015$. Da 0,8997 er nærmest 0,9, vælger vi $\Phi^{-1}(0,9) = 1,28$. Med andre ord er sandsynligheden (omtrent) 80% for at en normalfordelt stokastisk variabel afviger fra middelværdien med højst 1,28 standardafvigelser. Dette stemmer fint overens med højre graf i figur 2.5. ◊

Det er ikke nødvendigt at anvende tabelopslag af Φ. Mange programmer til matematiske og statistiske beregninger har dem indbygget. Et eksempel er Matlab:

Eksempel 13 (Matlab)
Hvis X er normalfordelt med middelværdi `mu` og standardafvigelse `sigma`, kan man finde $P(X \leq \mathtt{x})$ i `Matlab` vha. kommandoen `normcdf(x,mu,sigma)`. Antag, at X følger en standard normalfordeling, dvs. $X \sim \mathcal{N}(0,1)$.

Bestem følgende sandsynlighed $P(X \leq 1,4)$.
Løsning: Ifølge tabel 6.1 er $P(X \leq 1,4) = 0,9192$. I Matlab får vi (ikke overraskende) det samme svar:

```
>> normcdf(1.4,0,1)
ans =
    0.9192
```

Den inverse fordelingsfunktion er implementeret i `Matlab` som `norminv`. Find x så $P(X \leq x) = 0,7$.
Løsning: Ifølge tabel 6.1 har vi $P(X \leq 0,52) = 0,6985$ og $P(X \leq 0,53) = 0,7019$, dvs. x ligger et sted mellem 0,52 og 0,53. Vha. Matlab finder vi, at svaret er $x = 0,5244$:

```
>> norminv(0.7,0,1)
ans =
    0.5244
```

Sætningen 7 siger, at en lineær transformation af en normalfordelt stokastisk variabel er normalfordelt. Generelt er en linearkombination af flere normalfordelte stokastiske variable normalfordelt:

Sætning 9
Antag, at $X_1, X_2, \ldots, X_n$ er normalfordelte stokastiske variable og $a_0, a_1, a_2, \ldots, a_n$ er reelle konstanter, da er linearkombinationen $a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$ også normalfordelt.

Øvelse 2
Antag at X er standard normalfordelt, dvs. $X \sim \mathcal{N}(0,1)$. Bestem følgende sandsynligheder

1. $P(X \leq 0,6)$
2. $P(X \geq -1,5)$
3. $P(-1 \leq X \leq 2)$

Øvelse 3
Antag at højden blandt værnepligtige mænd kan betragtes som værende normalfordelt med middelværdi 173,3 cm og varians 6,4^{2}cm^2. Antag vi har en stikprøve på 1000 værnepligtige mænd. Af disse 1000, hvor mange vil vi (ca.) forvente er...

1. *...over 170cm højde?*
2. *...over 180cm højde?*
3. *...over 190cm højde?*
4. *...over 200cm højde?*

Øvelse 4
Antag, at $X \sim \mathcal{N}(\mu, \sigma^2)$. Bestem følgende sandsynligheder

1. $P(\mu - \sigma \leq X \leq \mu + \sigma)$
2. $P(2\mu - \sigma \leq X \leq \mu + 2\sigma)$
3. $P(3\mu - \sigma \leq X \leq \mu + 3\sigma)$

Øvelse 5
Antag, at vægten, en kylling øges med over en uge, er normalfordelt med middelværdi 350g og varians 30 g^2.

1. *Hvad er sandsynligheden for en tilvækst på mere end 300g?*

Antag, at vi har målt vægtforøgelsen på to kyllinger. Antag desuden, at vægtforøgelserne på de to kyllinger er uafhængige af hinanden.

2. *Hvad er sandsynligheden for, at begge kyllinger har en vægtforøgelse mindre end 300g?*

3. *Hvad er sandsynligheden for, at mindst en af de to kyllinger har en vægtforøgelse på mere end 300g?*

2.5 Tilfældig fejl

Som nævnt i indledningen til dette kapitel er formålet med dette kapitel at formulere en matematisk model for måling af fx. vinkler. Lad μ betegne en *sand* vinkel i trekanten i figur 1.1, og lad X være en måling af denne vinkel. I almindelighed vil målingen X afvige fra μ. Vi betegner fejlen ϵ, dvs.

$$X = \mu + \epsilon.$$

Vi betragter ϵ som en tilfældig fejl, og vi vil derfor tænke på ϵ som en stokastisk variabel. Specifikt vil vi antage, at

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

dvs. fejlen ϵ er normalfordelt med middelværdi nul og varians σ^2. Jf. sætning 2 og 4, har vi

$$\mathbb{E}[X] = \mathbb{E}[\mu + \epsilon] = \mathbb{E}[\mu] + \mathbb{E}[\epsilon] = \mu + 0.$$

og

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{V}\mathrm{ar}[\mu + \epsilon] = \mathbb{V}\mathrm{ar}[\epsilon] = \sigma^2.$$

Dvs. at i middel er vores måling X lig den sande vinkel μ. Af sætning 7 følger det, at målingen X desuden er normalfordelt:

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

Fra teorien om normalfordelingen har vi, at 95% af målingerne vil ligge i intervallet $\mu \pm 1,96\sigma$. I udgangspunktet kender vi ikke σ^2.

2.6 Linearisering

Vi har i det forrige afsnit set på, hvordan man finder middelværdi og varians for en lineær transformation af en stokastisk variabel. Antag, at vi for trekanten i figur 1.1 har målt vinklen α og sidelængderne S_b og S_c. Man kan da udregne arealet vha. $\frac{1}{2}\sin(\alpha)S_bS_c$. I praksis har vi en ide om, hvor usikre vores målinger af vinkler og afstande er, men hvordan påvirker det usikkerheden på arealet det beregnede areal? Vi starter med et lidt simplere eksempel.

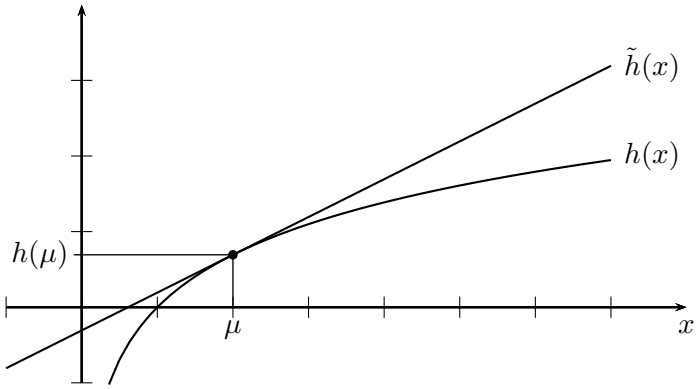
Antag, at $X \sim \mathcal{N}(\mu, \sigma^2)$, dvs. X er en normalfordelt stokastisk variabel med middelværdi μ og varians σ^2. Definer en ny stokastisk variabel $Y = h(X)$, hvor h er en differentiabel funktion. Med mindre $h(x)$ er en lineær funktion, kan vi ikke umiddelbart sige, hvilken fordeling Y følger — eller bare, hvilken middelværdi og varians Y har. Løsningen er at approksimere $h(x)$ med en lineær funktion. Denne fremgangsmåde kaldes, at man *lineariserer* $h(x)$. I det følgende betegner $h'(x)$ den afledte af $h(x)$ mht. x, dvs. $h'(x) = \mathrm{d}h(x)/\mathrm{d}x$.

Lineariseringen består i at approksimere $h(x)$ med en funktion, der svarer til tangentlinjen til funktionen $h(x)$ i punktet $(\mu, h(\mu))$. Linjen har hældning $h'(\mu) = \mathrm{d}h(x)/\mathrm{d}x|_{x=\mu}$ og går igennem punktet $(\mu, h(\mu))$. Det er let at vise, at forskriften for linjen er

$$\begin{aligned}\tilde{h}(x) &= h(\mu) + (x - \mu)h'(\mu) \\ &= h(\mu) - \mu h'(\mu) + xh'(\mu).\end{aligned}$$

Eksempel 14
Figur 2.7 viser et eksempel på linearisering, hvor $h(x) = \ln(x)$ og $\mu = 2$. I dette tilfælde er $h'(x) = 1/x$, og $h'(\mu) = 1/2$. Lineariseringen af $h(x)$ er derfor

$$\begin{aligned}\tilde{h}(x) &= \ln(2) - 2\frac{1}{2} + x\frac{1}{2} \\ &= \ln(2) - 1 + \frac{1}{2}x.\end{aligned}$$



Figur 2.7: Et eksempel på linearisering, hvor $h(x) = \ln(x)$ og $\mu = 2$.

I dette eksempel er det lidt problematisk at antage, at X er normalfordelt, hvorfor? $\Diamond$

Middelværdien og variansen for $Y = h(X)$ kan nu approksimeres ved

$$\mathbb{E}[Y] \approx h(\mu) \quad \text{og} \quad \mathbb{V}\text{ar}[Y] \approx (h'(\mu))^2\sigma^2.$$

Approksimationen af middelværdien $\mathbb{E}[Y]$ følger af:

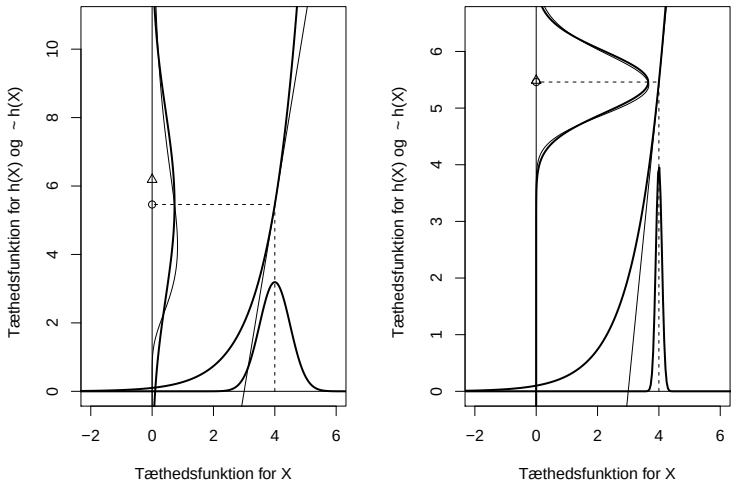
$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[h(X)] \\ &\approx \mathbb{E}[h(\mu) - \mu h'(\mu) + X h'(\mu)] \\ &= h(\mu) - \mu h'(\mu) + \mu h'(\mu) = h(\mu)\end{aligned}$$

og approksimationen for variansen $\mathbb{V}\text{ar}[Y]$ kan udledes vha.

$$\mathbb{V}\text{ar}[Y] = \mathbb{V}\text{ar}[h(X)] \approx \mathbb{V}\text{ar}[h(\mu) - \mu h'(\mu) + X h'(\mu)] = (h'(\mu))^2\sigma^2.$$

Lineariseringen $\tilde{h}(x)$ er en god approksimation af $h(x)$ så længe, at x er tæt på μ. Derfor er approksimationerne af $\mathbb{E}[Y]$ og $\mathbb{V}\text{ar}[Y]$ også kun gode, hvis X med stor sandsynlighed er tæt på μ. Dette er tilfældet, hvis X's varians ikke er for stor. Det illustreres i næste eksempel.

Eksempel 15
Antag, at $X \sim \mathcal{N}(\mu, \sigma^2)$, og $h(x) = \exp(x)$. I dette tilfælde er $h'(x) = \exp(x)$ og lineariseringen af $h(x)$ er $\tilde{h}(x) = \exp(\mu) - \mu\exp(\mu) + x\exp(\mu)$. $\mathbb{E}[Y] \approx \exp(\mu)$ og $\mathbb{V}\text{ar}[Y] \approx \exp(\mu)^2\sigma^2 = \exp(2\mu)\sigma^2$. Figur 2.8 viser situationen for to forskellige valg af standardafvigelse σ. I det venstre plot i figur 2.8 er $\sigma^2 = 0,5$. I det højre plot er $\sigma = 0,1$. I begge tilfælde er $\mu = 4$. I det venstre plot er der en tydelig forskel mellem den sande tæthedsfunktion (tynd linje) og tæthedsfunktionen, der er et resultat af lineariseringen (tyk linje).



Figur 2.8: Linearisering af log-normalfordelte stokastiske variable.

Man kan vise, at den *sande* middelværdi og varians er $\mathbb{E}[Y] = \exp(\mu+\sigma^2/2)$ og $\mathbb{V}\mathrm{ar}[Y] = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. Hvis variansen σ^2 er lille i forhold til middelværdien μ, er de tilnærmede udtryk ovenfor tæt på de sande værdier.

2.7 Kovarians

Indtil nu har vi beskæftiget os med uafhængige stokastiske variable. Fra sætning 3 ved vi, at udregning af middelværdien for en linearkombination af stokastiske variable ikke forudsætter uafhængighed mellem de enkelte stokastiske variable. I sætning 5 er det derimod en forudsætning, at de stokastiske variable er indbyrdes uafhængige. I dette afsnit vil vi indføre begrebet kovarians for at kunne håndtere variansen af en linearkombination af indbyrdes *afhængige* stokastiske variable.

Definition 10 (Kovarians)
Antag, at X og Y er to stokastiske variabel med middelværdier μ_X og μ_Y. Kovariansen mellem de stokastiske variable X og Y er da defineret som

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

Kovariansen kan tage både negative og positive værdier. Kovariansen er positiv, hvis store værdier af X generelt følges med store værdier af Y. Ligedes er kovariansen negativ, hvis store værdier af X generelt følges med små værdier af Y. Her skal stor værdi af X forstås som en værdi større end μ_X. Tilsvarende er en lille værdi af X en værdi mindre end μ_X.

Bemærk, at $\mathrm{Cov}(X,X) = \mathbb{V}\mathrm{ar}[X]$ idet $\mathrm{Cov}(X,X) = \mathbb{E}[(X-\mu_X)(X-\mu_X)] = \mathbb{E}[(X-\mu_X)^2] = \mathbb{V}\mathrm{ar}[X]$, hvor det sidste lighedstegn følger af definitionen på varians.

Eksempel 16
Fridjof oplyser, at kovariansen mellem antal solgte æbler og pærer er 139,5. Det tyder på, at antallet af solgte æbler og pærer følges ad. Hvor stærk denne sammenhæng er, er svært at sige, men det vender vi tilbage til. ◊

Sætning 10 (Kovarians og uafhængighed)
Hvis X og Y er uafhængige stokastiske variable, er $\mathrm{Cov}(X,Y) = 0$. *Det modsatte gælder generelt ikke. Dvs. hvis* $\mathrm{Cov}(X,Y) = 0$ *kan vi ikke konkludere, at X og Y er uafhængige.*

Vha. kovarianser er det muligt at udregne variansen af linearkombinationer af stokastiske variable uden at antage uafhængighed. Vi starter med variansen for summen af to stokastiske variable:

Sætning 11
Summen af to stokastiske variable X og Y har variansen

$$\mathbb{V}\mathrm{ar}[X+Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\mathrm{Cov}(X,Y).$$

Bevis:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X+Y] &= \mathbb{E}[(X+Y-E[X+Y])^2] \\
&= \mathbb{E}[(X+Y-(\mu_X+\mu_Y))^2] \\
&= \mathbb{E}[((X-\mu_X)+(Y-\mu_Y))^2] \\
&= \mathbb{E}[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{E}[(X-\mu_X)^2] + \mathbb{E}[(Y-\mu_Y)^2] + 2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\mathrm{Cov}(X,Y).
\end{aligned}$$

Eksempel 17
Fridjof fra før oplyser, at variansen for antal solgte æbler og pærer er hhv. 121,4 og 327,2. Hvad er variansen af det samlede antal solgte stykker frugt?
Løsning: Vi definerer X og Y som i det første eksempel og lader M betegne det samlede antal solgte stykker frugt, dvs. $M = X+Y$. Variansen for M finder vi ved:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[M] &= \mathbb{V}\mathrm{ar}[X+Y] \\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\mathrm{Cov}(X,Y) \\
&= 121,4 + 327,2 + 2 \cdot 139,5 \\
&= 727,6.
\end{aligned}$$

Bemærk, at variansen af summen (727,6) er meget større end summen af de to varianser (121,4 + 327,2 = 448,6). ◊

Regnereglen for variansen af summen af to stokastiske variable kan udvides til en generel linearkombination af to stokastiske variable:

Sætning 12
Variansen af linearkombinationen $aX + bY + c$ af de stokastiske variable X og Y er givet ved

$$\mathbb{V}\text{ar}[aX + bY + c] = a^2\mathbb{V}\text{ar}[X] + b^2\mathbb{V}\text{ar}[Y] + 2ab\text{Cov}(X, Y).$$

Bevis Beviset tager udgangspunkt i definitionen 7 for varians:

$$\begin{aligned}
\mathbb{V}\text{ar}[aX + bY + c] &= \mathbb{E}[(aX + bY + c - \mathbb{E}[aX + bY - c])^2] \\
&= \mathbb{E}[(aX + bY + c - (a\mu_X + b\mu_Y - c))^2] \\
&= \mathbb{E}[(a(X - \mu_X) + b(Y - \mu_Y))^2] \\
&= \mathbb{E}[a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)] \\
&= a^2\mathbb{E}[(X - \mu_X)^2] + b^2\mathbb{E}[(Y - \mu_Y)^2] + 2ab\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= a^2\mathbb{V}\text{ar}[X] + b^2\mathbb{V}\text{ar}[Y] + 2ab\text{Cov}(X, Y),
\end{aligned}$$

hvor sidste lighed følger af definitionen 7 for varians og definition 10 for kovarians.□

Eksempel 18
Hvad er variansen af Fridjofs daglige overskud?
Løsning: Husk, at overskuddet er givet ved $S = 1,27X + 0,87Y - 119$. Vi kan nu finde variansen for S:

$$\begin{aligned}
\mathbb{V}\text{ar}[S] &= \mathbb{V}\text{ar}[1,27X + 0,87Y - 119] \\
&= 1,27^2\mathbb{V}\text{ar}[X] + 0,87^2\mathbb{V}\text{ar}[Y] + 2 \cdot 1,27 \cdot 0,87\text{Cov}(X, Y) \\
&= 1,27^2 \cdot 121,4 + 0,87^2 \cdot 327,2 + 2 \cdot 1,27 \cdot 0,87 \cdot 139,5 \\
&= 751,73.
\end{aligned}$$

Fortolkning: Hvis vi antager, at overskuddet er normalfordelt, så vil 95% af alle dage generere et overskud i intervallet $\mathbb{E}[S] \pm 1,96\sqrt{\mathbb{V}\text{ar}[S]} = 59,729 \pm 1,96\sqrt{751,73} \approx 59,729 \pm 53,74 \approx [5,88; 113,47]$. Dvs. de fleste dage ligger Fridjofs overskud mellem ca. 6kr og 113 kr. Hvor urimelig er normalfordelingsantagelsen? ◊

Bemærk at variansen for differencen $X - Y$ er givet ved

$$\begin{aligned}
\mathbb{V}\text{ar}[X - Y] &= \mathbb{V}\text{ar}[X] + (-1)^2\mathbb{V}\text{ar}[Y] + 1 \cdot (-1)2\text{Cov}(X, Y) \\
&= \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y] - 2\text{Cov}(X, Y).
\end{aligned}$$

Sætning 12 kan udvides til et udtryk for variansen af en linearkombination af et vilkårligt antal stokastiske variable:

Sætning 13 (Varians for linearkombinationer)
Antag $X_1, \ldots, X_n$ er n stokastiske variable og $a_0, a_1, \ldots, a_n$ er reelle konstanter. Da er variansen af linearkombinationen $a_0 + a_1X_1 + \ldots + a_nX_n$ givet ved

$$\mathbb{V}\text{ar}[a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i^2 \mathbb{V}\text{ar}[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \text{Cov}(X_i, X_j). \quad (2.14)$$

Det er ofte bekvemt at benytte følgende notation for varians og kovarians: $\mathbb{V}\text{ar}[X_i] = \sigma_i^2$ og $\text{Cov}(X_i, X_j) = \sigma_{ij}$. Bemærk, at $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, dvs. $\sigma_{ij} = \sigma_{ji}$. Med denne notation kan (2.14) omskrives til

$$\mathbb{V}\text{ar}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i<j} a_i a_j \sigma_{ij}, \quad (2.15)$$

hvor $\sum_{i<j}$ betegner en sum over alle kombinationer af i og j, hvor i er (skarpt) mindre end j.

2.8 Korrelation

Kovariansen kan være svær at fortolke, da den afhænger af den enhed X og Y er målt i. Hvis X og Y er fortjenesten på hhv. æbler og pærer, så vil størrelsen af $\text{Cov}(X, Y)$ afhænge af, hvilken valuta de to fortjenester opgøres i. For at afhjælpe dette problem indfører vi korrelationen.

Definition 11 (Korrelation)
Korrelationen mellem to stokastiske variable X og Y betegnes $\text{Corr}(X, Y)$, og er defineret som

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\mathbb{V}\text{ar}[X]\mathbb{V}\text{ar}[Y]}}.$$

Korrelationen er et tal mellem -1 og 1 og kan ses som et udtryk for graden af lineær sammenhæng. Hvis $|\text{Corr}(X, Y)| = 1$ (den numeriske værdi af korrelationen er 1), er der perfekt lineær sammenhæng mellem X og Y. Det betyder, at $Y = aX + b$ for reelle tal a og b, se også Eksempel 20 nedenfor. Hvis X og Y er uafhængige er $\text{Corr}(X, Y) = 0$. Husk, at $\text{Corr}(X, Y) = 0$ ikke er ensbetydende med uafhængighed. Korrelationen mellem to stokastiske variable betegnes ofte ved det græske bogstav ρ [rho]. Dvs. $\rho = \text{Corr}(X, Y)$.

Eksempel 19
Antag vi har to stokastiske variable X og Y og to reelle positive konstanter $a > 0$ og $b > 0$. Hvad er da korrelationen mellem aX og bY?
Løsning: Vi starter med at bemærke, at

$$\begin{aligned}\mathbb{V}\text{ar}[aX] &= a^2\mathbb{V}\text{ar}[X] \\ \mathbb{V}\text{ar}[bY] &= b^2\mathbb{V}\text{ar}[Y].\end{aligned}$$

Yderligere gælder $\text{Cov}(aX, bY) = ab\text{Cov}(X, y)$. Kombineres disse resultater med Definition 11 på kovarians fås

$$\text{Corr}(aX, bY) = \text{Corr}(X, Y)$$

Eksempel 19 viser, at en skalering af X og Y ikke påvirker korrelationen. Hvis vi tænker på a og b som konverteringsfaktorer, når vi skifter valuta, understreger denne regneregel, at valg af valuta er irrelevant, når vi udregner korrelationen mellem fortjenesten på æbler og pærer.

Eksempel 20
Antag, at X er en stokastisk variabel og $Y = aX + b$, $a \neq 0$. Hvad er da korrelationen mellem X og Y? For at finde korrelationen skal vi først finde variansen for Y samt kovariansen mellem X og Y.

Det følger af sætning 4, at

$$\mathbb{V}\text{ar}[Y] = a^2\mathbb{V}\text{ar}[X].$$

Kovariansen mellem X og Y er

$$\begin{aligned}\text{Cov}(X, Y) &= \text{Cov}(X, aX + b)\\ &= \mathbb{E}[(X - \mathbb{E}[X])(aX + b - a\mathbb{E}[X] - b)]\\ &= \mathbb{E}[a(X - \mathbb{E}[X])(X - \mathbb{E}[X])]\\ &= a\mathbb{V}\text{ar}[X].\end{aligned}$$

Korrelationen er da givet ved

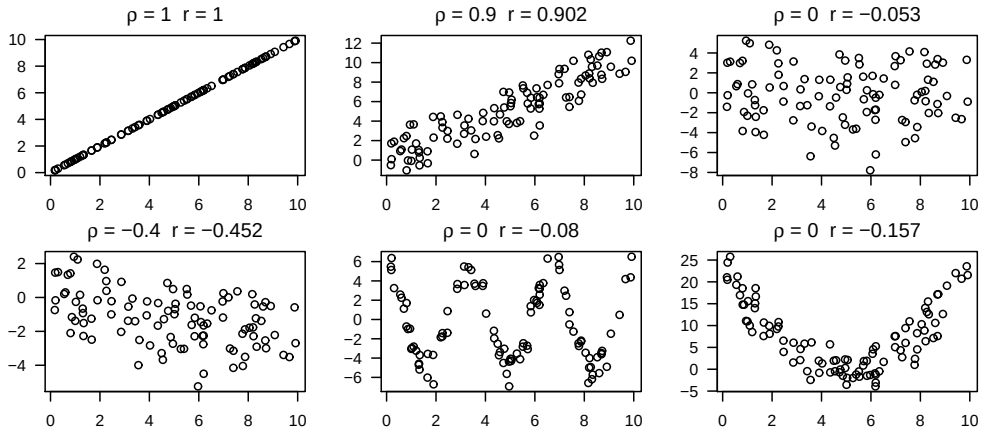
$$\begin{aligned}\text{Corr}(X, Y) &= \text{Cov}(X, Y)/\sqrt{\mathbb{V}\text{ar}[X]\mathbb{V}\text{ar}[Y]}\\ &= a\mathbb{V}\text{ar}[X]/\sqrt{\mathbb{V}\text{ar}[X]a^2\mathbb{V}\text{ar}[X]}\\ &= a/\sqrt{a^2}.\end{aligned}$$

Hvis $a > 0$, får vi $\text{Corr}(X, Y) = 1$, og hvis $a < 0$, får vi $\text{Corr}(X, Y) = -1$. Dvs. der er en perfekt lineær sammenhæng mellem X og Y. ◊

Eksempel 21
Find korrelationen mellem antallet af solgte æbler og pærer.
Løsning: Korrelationen er $\rho = \text{Cov}(X, Y)/\sqrt{\mathbb{V}\text{ar}[X]\mathbb{V}\text{ar}[Y]} = 139,5/\sqrt{121,4 \cdot 327,2} \approx 0,70$. Med en korrelation på 0,7 er der en tydelig sammenhæng mellem salget af æbler og pærer. ◊

2.9 Matrix formulering

Når der er mange stokastiske variable i sving samtidigt, kan det hjælpe at reformulere problemet i termer af vektorer og matricer.



Figur 2.9: Seks plots af stikprøver fra populationer, hvor korrelationen er ρ som angivet over hvert plot. Over hvert plot er desuden angivet stikprøvekorrelationen r, der er et estimat af ρ, se afsnit 3.2.

Hvis $X_1, \ldots, X_n$ er stokastiske variable, da er

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

en (n dimensional) stokastisk vektor. Antag, at X_i har middelværdi $\mathbb{E}[X_i] = \mu_i$ og varians $\mathbb{V}\text{ar}[X_i] = \sigma_i^2$. Da er middelværdien for den stokastiske vektor $\mathbf{X}$ givet ved

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \boldsymbol{\mu}.$$

Definition 12 (Kovariansmatrix)
Variansen for $\mathbf{X}$ er en $n \times n$ matrix, der betegnes kovariansmatricen og er givet som

$$\begin{aligned} \mathbb{V}\text{ar}[\mathbf{X}] &= \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu})\right] \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} \\ &= \mathbf{K}_{\mathbf{X}}, \end{aligned}$$

hvor vi, som i sætning 13, har, at $\sigma_i^2 = \mathbb{V}\text{ar}[X_i]$ og $\sigma_{ij} = \text{Cov}(X_i, X_j)$, $i \neq j$. $\triangle$

Bemærk, at kovariansmatricen $\mathbf{K_X}$ er symmetrisk da $\sigma_{ij} = \sigma_{ji}$ og $\mathbf{K_X}$ har varianserne på diagonalen.

Eksempel 22
Antag X_1 og X_2 er to stokastiske variable, hvor middelværdierne er hhv. $\mathbb{E}[X_1] = 1.7$ og $\mathbb{E}[X_2] = 2.3$, varianserne er $\mathbb{V}\text{ar}[X_1] = 7.2$ og $\mathbb{V}\text{ar}[X_2] = 6.4$, og sluttelig er kovariansen mellem de to stokastiske variable $\text{Cov}(X_1, X_2) = 5.2$. I dette tilfælde er middelværdivektoren og kovariansmatricen givet ved

$$\boldsymbol{\mu} = \begin{bmatrix} 1.7 \\ 2.3 \end{bmatrix} \quad \text{og} \quad \mathbf{K_X} = \begin{bmatrix} 7.2 & 5.2 \\ 5.2 & 6.4 \end{bmatrix}.$$

Sætningerne 2 og 4 angiver regneregler for middelværdi og varians for en lineær transformation $aX + b$ af én stokastisk variabel X. Følgende sætning opsummerer tilsvarende regneregler for en stokastisk vektor.

Sætning 14
Antag, at $\mathbf{X}$ *er en* n *dimensional stokastisk (søjle)vektor med middelværdi* $\boldsymbol{\mu}$ *og kovariansmatrix* $\mathbf{K_X}$*. Lad* $\mathbf{A}$ *være en vilkårlig* $m \times n$ *matrix og* $\mathbf{b}$ *en vilkårlig* m *dimensional søjlevektor. Middelværdien af den lineære transformation* $\mathbf{AX} + \mathbf{b}$ *er*

$$\mathbb{E}[\mathbf{AX} + \mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}.$$

På tilsvarende vis er kovariansen for den lineære transformation $\mathbf{AX} + \mathbf{b}$ *givet ved*

$$\mathbb{V}\text{ar}[\mathbf{AX} + \mathbf{b}] = \mathbf{A}\mathbb{V}\text{ar}[\mathbf{X}]\mathbf{A}^T = \mathbf{A}\mathbf{K_X}\mathbf{A}^T.$$

Eksempel 23
Antag, at X_1 og X_2 er givet som i eksempel 22. Antag desuden, at

$$\mathbf{A} = \begin{bmatrix} 1.7 & 2.1 \\ 3.4 & 4.2 \\ 5.1 & 6.9 \end{bmatrix} \quad \text{og} \quad \mathbf{b} = \begin{bmatrix} 9.2 \\ 8.7 \\ 7.4 \end{bmatrix}.$$

Da er $\mathbb{E}[\mathbf{X}]$ og $\mathbb{V}\text{ar}[\mathbf{X}]$ givet ved

$$\begin{aligned} \mathbb{E}[\mathbf{AX} + \mathbf{b}] &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\ &= \begin{bmatrix} 1.7 & 2.1 \\ 3.4 & 4.2 \\ 5.1 & 6.9 \end{bmatrix} \begin{bmatrix} 1.7 \\ 2.3 \end{bmatrix} + \begin{bmatrix} 9.2 \\ 8.7 \\ 7.4 \end{bmatrix} = \begin{bmatrix} 16.95 \\ 24.14 \\ 31.94 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \mathbb{V}\text{ar}[\mathbf{AX} + \mathbf{b}] &= \mathbf{A}\mathbf{K_X}\mathbf{A}^T \\ &= \begin{bmatrix} 1.7 & 2.1 \\ 3.4 & 4.2 \\ 5.1 & 6.9 \end{bmatrix} \begin{bmatrix} 7.2 & 5.2 \\ 5.2 & 6.4 \end{bmatrix} \begin{bmatrix} 1.7 & 3.4 & 5.1 \\ 2.1 & 4.2 & 6.9 \end{bmatrix} \\ &= \begin{bmatrix} 86.160 & 172.320 & 271.848 \\ 172.320 & 344.640 & 543.696 \\ 271.848 & 543.696 & 857.952 \end{bmatrix}. \end{aligned} \tag{2.16}$$

Definer en stokastisk vektor

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix},$$

som er givet ved $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, hvor $\mathbf{A}$, $\mathbf{X}$ og $\mathbf{b}$ er som givet ovenfor. Varianser og kovarianser for Y_1, Y_2 og Y_3 kan nu aflæses i kovariansmatricen (2.16). Fx. er variansen $\mathbb{V}\mathrm{ar}[Y_1] = 86,160$ og kovariansen mellem Y_2 og Y_3 er $\mathrm{Cov}(Y_2, Y_3) = 543.696$. Korrelationen mellem Y_2 og Y_3 er

$$\frac{\mathrm{Cov}(Y_2, Y_3)}{\sqrt{\mathbb{V}\mathrm{ar}[Y_2]\mathbb{V}\mathrm{ar}[Y_3]}} = \frac{543.696}{\sqrt{344.640 \cdot 857.952}} = 0.999866,$$

dvs. Y_2 og Y_3 er næsten perfekt korrelerede.

Kapitel 3

Estimation

Ved hjælp af en teodolit måles en bestemt vinkel n gange. Vinklens sande værdi er μ gon. Vinkelmålingerne foregår uafhængigt af hinanden og under samme omstændigheder. Det antages videre, at de tilfældige målefejl, der begås, følger en normalfordeling. Målingerne resulterer i et datamateriale:

$$x_1, \ldots, x_n.$$

Eksempel 24
En vinklen er målt med 10 satser. Følgende værdier er observeret:

$$\begin{aligned}
x_1 &= 164{,}508 \text{ gon} \\
x_2 &= \quad ,509 \quad \text{”} \\
x_3 &= \quad ,511 \quad \text{”} \\
x_4 &= \quad ,507 \quad \text{”} \\
x_5 &= \quad ,510 \quad \text{”} \\
x_6 &= \quad ,511 \quad \text{”} \\
x_7 &= \quad ,517 \quad \text{”} \\
x_8 &= \quad ,510 \quad \text{”} \\
x_9 &= \quad ,514 \quad \text{”} \\
x_{10} &= \quad ,513 \quad \text{”}
\end{aligned}$$

Spørgsmålet er nu, hvad er vores bedste bud på den sande vinkel, og hvilken usikkerhed er der forbundet med dette bud. ◊

Med henblik på at gennemføre en detaljeret analyse af det foreliggende datamateriale formuleres en statistisk model for, hvordan vi mener, data er fremkommet. Det antages således, at målingerne $x_1, \ldots, x_n$ er realisationer af uafhængige stokastiske variable $X_1, \ldots, X_n$, alle med samme middelværdi μ og varians σ^2. Ofte antages også, at de stokastiske variable er normalfordelte, dvs. $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Variansen σ^2 er et mål for nøjagtigheden af målingerne.

Det er nu opgaven ved hjælp af den observerede stikprøve $x_1, \ldots, x_n$ at skønne - *estimere* - den givne normalfordelings middelværdi μ (dvs. den sande vinkel) og samme fordelings varians σ^2 (dvs. målet for målenøjagtigheden).

3.1 Estimation af middelværdi og varians

Til at estimere μ benyttes almindeligvis stikprøvegennemsnittet:

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.1}$$

Til at estimere σ^2 anvendes normalt stikprøvevariansen:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.2}$$

På baggrund af vores stikprøve får vi:

$$\bar{x} = 164,511 \text{ gon} \quad \text{og} \quad s^2 = (0,003)^2\text{gon}^2$$

Det forekommer umiddelbart mærkeligt, at man ved definitionen af s^2 anvender faktoren $\frac{1}{n-1}$ og ikke $\frac{1}{n}$. En forklaring herpå gives i Sætning 17 i det følgende.

Vi skal herefter ved hjælp af den formulerede statistiske model indse, at estimaterne $\bar{x}$ og s^2 er gode estimater for hhv. μ og σ^2. Vi vil desuden angive en metode til at afgøre, hvor nøjagtige estimaterne $\bar{x}$ og s^2 er. Bemærk, at ligesom der til hver observation x_i svarer en stokastisk variabel X_i, så svarer der til estimaterne $\bar{x}$ og s^2 stokastiske variable $\bar{X}$ og S^2 som illustreret i følgende diagram:

$$\begin{array}{ccccc} X_1, \ldots, X_n & & \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i & & S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)^2 \\ \downarrow \quad \downarrow & & \downarrow & & \downarrow \\ x_1, \ldots, x_n & & \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i & & s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{array} \tag{3.3}$$

Vi benytter i det følgende betegnelsen *estimator* for den stokastiske variabel svarende til et estimat. Dvs. $\bar{X}$ er eksempelvis estimatoren svarende til estimatet $\bar{x}$.

Sætning 15
Antag $X_1, \ldots, X_n$ er en stikprøve fra en fordeling, der har middelværdi μ og varians σ^2. Da gælder

$$\mathbb{E}[\bar{X}] = \mu. \tag{3.4}$$

Hvis yderligere $X_1, \ldots, X_n$ er uafhængige gælder

$$\mathbb{V}\text{ar}[\bar{X}] = \frac{\sigma^2}{n}. \tag{3.5}$$

Hvis $X_1, \ldots, X_n$ desuden er normalfordelte $\mathcal{N}(\mu, \sigma^2)$, da er $\bar{X}$ normalfordelt $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Bevis Vi starter med at bemærke, at gennemsnittet kan skrives som $\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$. Gennemsnittet er således en linearkombination af stokastiske variable. Det følger derfor af sætning 3 og sætning 5, at

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu$$

og

$$\mathbb{V}\mathrm{ar}[\bar{X}] = \mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}[X_i] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Her har vi desuden brugt, at $X_1, \ldots, X_n$ har samme middelværdi og varians, og at $X_1, \ldots, X_n$ er uafhængige.

At stikprøvegennemsnittet er normalfordelt, når stikprøven er fra en normalfordeling, er en konsekvens af sætning 9. □

Hvis vores stikprøve er fra en normalfordeling, siger sætning 15, at $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Fra afsnittet om normalfordelingen ved vi, at $\bar{X}$ med 95% sandsynlighed vil ligge i intervallet $\mu \pm 1.96\sigma/\sqrt{n}$. Jo større stikprøve (dvs. jo større n) jo oftere vil $\bar{x}$ ligge tæt på n.

Et tilsvarende resultat for fordelingen af $\bar{X}$ gælder, selvom $X_1, \ldots, X_n$ ikke er normalfordelt, når blot n er tilstrækkelig stor:

Sætning 16 (Central grænseværdisætning)
Antag $X_1, \ldots, X_n$ er uafhængige og identisk fordelte stokastiske variable med middelværdi μ og varians σ^2. Da følger

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

approksimativt en normalfordeling med middelværdi 0 og varians 1. Jo større stikprøvestørrelse n, jo bedre er approksimationen. Mere præcist, antag $Z \sim \mathcal{N}(0,1)$ da gælder, at

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) = P(Z \leq x) \quad \textit{for alle } x \in \mathbf{R}.$$

Dvs. fordelingsfunktionen for $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ligner mere og mere fordelingsfunktionen for en standard normalfordeling.

Af den centrale grænseværdisætning fås, at hvis $X_1, \ldots, X_n$ er uafhængige og ensfordelte, og n er tilstrækkelig stor, så er gennemsnittet approksimativt normalfordelt $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Vi vender nu tilbage til det oprindelige eksempel vedrørende vinkelmåling. Da vi har antaget, at fejlene er normalfordelte, har vi, at $\bar{X}$ er normalfordelt $N(\mu, \frac{\sigma^2}{n})$.

Det betyder, at $\bar{X}$ er en stokastisk variabel, der har μ "centralt" placeret i sin sandsynlighedsfordeling, idet $\mathbb{E}[\bar{X}] = \mu$. Desuden er $\bar{X}$'s sandsynlighedmasse samlet omkring μ, idet $\mathbb{V}\text{ar}[\bar{X}] = \frac{\sigma^2}{n}$ er "lille". Vi kan altså forvente, at en observeret værdi af $\bar{X}$ ligger tæt på μ.

Bemærk: Estimatoren $\bar{X}$ kaldes en *central estimator* for μ, fordi $\mathbb{E}[\bar{X}] = \mu$. Tilsvarende kaldes $\bar{x}$ et *centralt estimat* for μ.

Analogt til sætning 15 kan vi vise, at S^2 er en central estimator for σ^2:

Sætning 17
Lad $X_1, \ldots, X_n$ være en stikprøve af uafhængige stokastiske variable med middelværdi μ og varians σ^2. Da gælder, at

$$\mathbb{E}[S^2] = \sigma^2 \tag{3.6}$$

dvs. S^2 er en central estimator for σ^2, og s^2 er et centralt estimat for σ^2. Desuden gælder der (i de tilfælde vi betragter), at

$$\mathbb{V}\text{ar}[s^2] \to 0 \text{ for } n \to \infty.$$

Bevis Vi viser kun den første del af sætningen.

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{1}^{n}X_i + \sum_{i=1}^{n}\bar{X}^2\right] \qquad (3.7)
\end{aligned}$$

Vi anvender nu følgende omskrivninger

$$\sum_{i=1}^{n} X_i = n\frac{1}{n}\sum_{i=1}^{n} X_i = n\bar{X},$$

og

$$\sum_{i=1}^{n} \bar{X}^2 = n\bar{X}^2.$$

Ved indsætning heraf i (3.7) fås

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right].
\end{aligned}$$

Hermed

$$\mathbb{E}(S^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}[\bar{X}^2]\right). \tag{3.8}$$

Fra omskrivningen (2.6) har vi

$$\sigma^2 = \mathbb{V}\mathrm{ar}(X_i) = \mathbb{E}(X_i^2) - (\mathbb{E}(X_i))^2 = \mathbb{E}(X_i^2) - \mu^2,$$

hvad der medfører:

$$\mathbb{E}(X_i^2) = \sigma^2 + \mu^2 \tag{3.9}$$

På tilsvarende vis finder vi følgende omskrivning:

$$\mathbb{E}[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2 \quad \text{(hvorfor?)} \tag{3.10}$$

Ved indsætning af (3.9) og (3.10) i (3.8) får vi

$$\begin{aligned}\mathbb{E}[S^2] &= \frac{1}{n-1}\left(\sum_{i=1}^{2}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)\\ &= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)\\ &= \sigma^2\end{aligned}$$

Dvs.

$$\mathbb{E}[S^2] = \sigma^2.$$

Dette afslutter beviset. □

Estimatoren S^2 er altså en stokastisk variabel med σ^2 "centralt"placeret i sin sandsynlighedsfordeling. Ydermere gælder det, at for n "stor", er sandsynlighedsmassen for S^2 "samlet"omkring σ^2 (fordi $\mathbb{V}\mathrm{ar}[S^2] \to 0$ for $n \to \infty$). Vi kan altså forvente, at en observeret værdi af S^2 er et godt estimat for σ^2.
Bemærkning: Hvis vi som estimator for σ^2 havde valgt:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

ville vi få

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] &= \mathbb{E}\left[\frac{1}{n}\cdot\frac{n-1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\\ &= \frac{n-1}{n}\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\\ &= \frac{n-1}{n}\sigma^2\end{aligned}$$

Dvs.

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Estimatoren $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ er hermed *ikke* en central estimator for σ^2.

Opsummering

Som estimat for den sande vinkel μ benyttes $\bar{x}$. Som estimat for σ^2 benyttes s^2. Begge estimater er centrale. Som estimat for spredningen σ på den enkelte vinkel benyttes $s = \sqrt{s^2}$. Som estimat for spredningen $\frac{\sigma}{\sqrt{n}}$ for gennemsnittet $\bar{X}$ benyttes $\frac{s}{\sqrt{n}}$.

For en god ordens skyld skal bemærkes, at ovenstående resultater vedrørende estimater for middelværdi μ og variansen σ^2 naturligvis også gælder, selvom der ikke er tale om måling af en vinkel. Det afgørende er, om den omtalte statistiske model (3.3) gælder, dvs. at data kan antages at være observationer af uafhængige identisk fordelte stokastiske variable.

Øvelse 6
Estimer middelværdi og spredning for vinkelmålingen, hvis resultater er givet i eksempel 24.

Foretag beregningen på grundlag af:

1. X_1, X_2 *og* X_3

2. X_1, X_2, X_3 *og* X_4

3. *tilføj* X_5 *osv.*

Bemærk variationen i $\bar{X}$ *og* s. *Hvornår virker estimaterne pålidelige?*

Spredningen estimeret på alle 10 observationer giver $s_v = 0,0030$ *gon* $\approx \sigma_v$ *(spredningen på den enkelte måling, dvs. spredningen på vinklen målt i 1 sats). Som bekendt dannes en vinkel som differens mellem to retninger. Benyt resultatet til at estimere spredningen på en retning målt med 1 sats (*$s_r \approx \sigma_r$*).*

Øvelse 7
Antag, at vi foretager n *uafhængige målinger (satser)* $X_1, \ldots, X_n$ *af vinklen* β, *og lad* $\bar{X}$ *betegne gennemsnittet af disse. Antag* $X_i \sim N(\beta, 5mgon^2)$, $i = 1, \ldots, n$. *Bestem* n, *således at*

$$P(|\bar{X} - \beta| < 3 \text{ mgon}) = 0.95$$

Fortolk resultaterne.

3.1.1 Estimation af varians: Kendt middelværdi

Ud over estimaterne nævnt ovenfor får vi i det følgende i ét tilfælde brug for estimatet $\hat{s}^2$, der er defineret ved

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.11}$$

Estimatet $\hat{s}^2$ anvendes til at estimere σ^2, når μ er kendt. (Bemærk: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ udnytter ikke kendskab til μ!).

Begrundelsen for at benytte $\hat{s}^2$ fremfor s^2 når μ er kendt fremgår af følgende sætning:

Sætning 18
Antag, at vi i n forsøg har observeret n identisk fordelte stokastiske variable $X_1, \ldots, X_n$, hver med varians σ^2. Da er

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

en central estimator for σ^2. Hvis $X_1, \ldots, X_n$ er normalfordelte gælder der yderligere,

$$\mathbb{V}\mathrm{ar}[\hat{S}^2] = 2\frac{\sigma^4}{n} < 2\frac{\sigma^4}{n-1} = \mathbb{V}\mathrm{ar}[S^2]$$

Bevis

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

Af definitionen på varians (2.5) følger, at

$$\mathbb{E}[(X_i - \mu)^2] = \sigma^2 \quad \text{for alle } i = 1, \ldots, n.$$

Dvs.

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

Sætningens anden påstand vises ikke. □

Er μ kendt foretrækker vi således $\hat{s}^2$ fremfor s^2, da $\hat{S}^2$ er en bedre estimator end S^2 idet $\mathbb{V}\mathrm{ar}[\hat{S}^2] < \mathbb{V}\mathrm{ar}[S^2]$.

3.2 Estimation af kovarians og korrelation

Antag vi har en stikprøve bestående af n par af observationer $(x_1, y_1), \ldots, (x_n, y_n)$. Som eksempel kunne x_i være antal solgte æbler den $i$te dag og y_i kunne være antallet af solgte pærer på den $i$te dag.

Definition: Stikprøvekovariansen for stikprøven $(x_1, y_1), \ldots, (x_n, y_n)$ betegnes s_{XY} og er givet ved

$$s_{XY} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

Definition: Stikprøvekorrelationen betegnes r og er givet ved

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

I figur 2.9 på side 29 ses seks eksempler på stikprøver af parvise observationer og de tilhørende stikprøvekorrelationer. Over hvert plot står stikprøvekorrelationen r sammen med den sande korrelation ρ. Bemærk, hvordan data med meget systematiske sammenhænge mellem x og y kan have en stikprøvekorrelation tæt på nul.

3.3 Konfidensinterval

Antag igen, at vi har observeret n stokastiske variable $X_1, \ldots, X_n$, der er normalfordelte $N(\mu, \sigma^2)$ og uafhængige. Det skal videre antages — for ikke at skabe unødvendige komplikationer — at variansen σ^2 er kendt. Ifølge sætning 15 gælder, at $\bar{X}$ normalfordelt med middelværdi μ og varians σ^2/n, i kort notation $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Fra afsnit 2.4 har vi derfor, at

$$P\left(\mu - 1,96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95. \tag{3.12}$$

Dvs. med 95% sandsynlighed ligger stikprøve-gennemsnittet $\bar{X}$ i intervallet $\mu \pm 1,96\frac{\sigma}{\sqrt{n}}$.

Bemærk, at uligheden

$$\mu - 1,96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq 1,96\frac{\sigma}{\sqrt{n}}$$

i (3.12) kan omskrives til

$$\bar{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}.$$

Sandsynligheden (3.12) kan derfor omskrives til

$$P\left(\bar{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95. \tag{3.13}$$

Man kan tillægge formel (3.13) følgende fortolkning. Sandsynligheden for, at $\bar{X}$ antager en værdi $\bar{x}$, således at intervallet givet ved $\bar{x} \pm 1,96\frac{\sigma}{\sqrt{n}}$ indeholder μ, er 95%.

Definition 13 (Konfidensinterval)
Intervallet

$$\left[\bar{X} - 1,96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}\right] \tag{3.14}$$

betegnes et 95% konfidensinterval for μ. De 95% omtales også som konfidensniveauet. △

Bemærk, at konfidensintervallet som udgangspunkt er et stokastisk interval, da gænserne for intervallet afhænger af den stokastiske variabel $\bar{X}$, og der er altså 95% sandsynlighed for, at intervallet indeholder μ.

Observeres en konkret værdi $\bar{x}$ af $\bar{X}$ fås et tilsvarende observeret konfidensinterval $\left[\bar{x} - 1,96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96\frac{\sigma}{\sqrt{n}}\right]$. Dette interval er deterministisk, dvs. der er kun to muligheder: enten er μ indenfor intervallet eller også er μ udenfor intervallet. For en given stikprøve kan vi ikke vide, hvilken af mulighederne der er tale om. Men hvis vi træffer valget altid at hævde, at det observerede konfidensinterval indeholder μ, så begår vi i snit kun en fejl i 5% af tilfældende: Antag vi observerer de n stokastiske variable k gange, dvs. vi får k observationsrækker med hver n tal. Hvermed fås k gennemsnit $\bar{x}_1, \ldots, \bar{x}_k$ og k konfidensintervaller hørende til de k middelværdier. For k stor kan vi da forvente, at 95% af intervallerne indeholder μ. Dette er grunden til, at vi i praksis benytter konfidensintervallet til at angive et interval af mulige μ-værdier i tilgift til det bedste bud $\bar{x}$. Bredden af konfidensintervallet angiver med hvor stor/lille sikkerhed vi kender μ.

På helt analog måde kan man konstruere konfidensintervaller med andre konfidensniveauer (fx. 99% ved at udskifte faktoren 1,96 med 2,58).

3.4 Konfidensinterval og linearisering

Antag $\hat{\mu}$ er et estimat af en ukendt størrelse μ, og at $\hat{\mu}$ er approksimativt $N(\mu, \sigma^2_{\hat{\mu}})$. Da følger af overvejelser som i forrige afsnit, at

$$[\hat{\mu} - 1.96\sigma_{\hat{\mu}}; \hat{\mu} + 1.96\sigma_{\hat{\mu}}]$$

er et approksimativt 95% konfidensinterval for μ.

Vi betragter nu situationen $\mu = h(\mu_X)$ hvor μ_X estimeres ved et gennemsnit $\bar{X}$ af n målinger $X_1, \ldots, X_n$ og $\bar{X} \sim N(\mu_X, \sigma^2_X/n)$. Vi bruger da estimatet $\hat{\mu} = h(\bar{X})$ og opnår vha. linearisering som i afsnit 2.6, at $\hat{\mu}$ er approximativt normalfordelt $N(\mu, \sigma^2_{\hat{\mu}})$ hvor $\sigma^2_{\hat{\mu}} = (h'(\mu_X))^2\sigma^2_X/n$. Dermed fås et approximativt 95% konfidensinterval

$$\left[h(\bar{X}) - 1,96\frac{|h'(\bar{X})|\sigma_X}{\sqrt{n}}; h(\bar{X}) + 1,96\frac{|h'(\bar{X})|\sigma_X}{\sqrt{n}}\right],$$

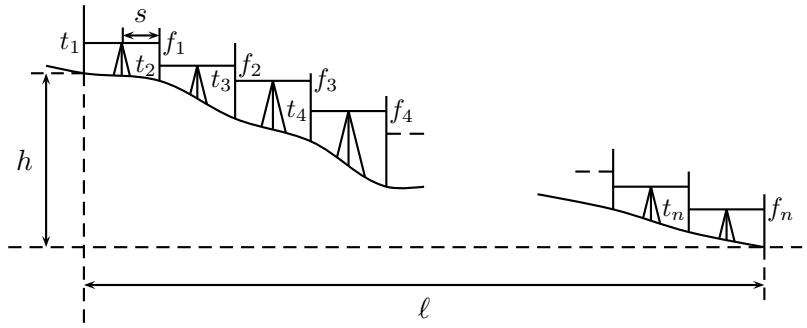
for μ, hvor $\sigma_{\hat{\mu}} = |h'(\bar{X})|\sigma_X/\sqrt{n}$ er en approksimation af $\hat{\mu}$s standardafvigelse. I praksis vil σ_X^2 ofte være ukendt og estimeres da ved $s_X^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$.

Kapitel 4

Fejlforplantning ved geometrisk nivellement

4.1 Geometrisk nivellement

Figur 4.1 illustrerer et geometrisk nivellement over en strækning ℓ bestående af n opstillinger.



Figur 4.1: Principskitse af et geometrisk nivellement over en strækning ℓ bestående af n opstillinger.

I den i'te opstilling foretages en stadieaflæsning ved tilbagesigte og fremsigte:

- t_i er stadieaflæsningen ved tilbagesigtet.
- f_i er stadieaflæsningen ved fremsigtet.

Højdeforskel h fremkommer som:

$$h = t_1 - f_1 + t_2 - f_2 + \ldots + t_n - f_n = [t - f] \tag{4.1}$$

Med "lige lange" sigter ($\sim s$) fås

$$l = 2ns. \tag{4.2}$$

Aflæsningerne på stadiet (t_i og f_i) forudsættes kun påvirket af tilfældige fejl, bl.a.:

- Fejl fra kompensator/libelle
- Aflæsningsfejl
- Fejl i stadiets inddelinger
- Refraktionsfejl

Stadieaflæsningerne t_i, f_i opfattes nu som værdier antaget af uafhængige stokastiske variable T_i, F_i med samme varians σ_a^2. $H = T_1 - F_1 + \ldots + T_n - F_n$ antager hermed værdien $h = t_1 - f_1 + \ldots + t_n - f_n$. Af (2.8) fås:

$$\mathbb{V}\mathrm{ar}[H] = \sigma_a^2 + \sigma_a^2 + \ldots + \sigma_a^2 + \sigma_a^2 = 2n \cdot \sigma_a^2. \tag{4.3}$$

Ifølge (4.2) har vi $2n = \frac{l}{s}$, hvorved

$$\mathbb{V}\mathrm{ar}[H] = \sigma_l^2 = \frac{\sigma_a^2}{s} \cdot l \quad \Rightarrow \quad \sigma_l = \frac{\sigma_a}{\sqrt{s}}\sqrt{l}. \tag{4.4}$$

Erfaringer viser, at størrelsen $\frac{\sigma_a}{\sqrt{s}}$ stort set kun afhænger af instrument og observationsmetoder, og kun i ringe grad varierer som funktion af s ($s < 100$ m). Den betragtes derfor som en karakteristisk konstant for observationsmetoden og benævnes σ_k. Indføres $\sigma_k = \frac{\sigma_a}{\sqrt{s}}$ fås af (4.4):

$$\sigma_l = \sigma_k\sqrt{l}. \tag{4.5}$$

Hermed er σ_l altså spredningen på et geometrisk nivellement over længden l.

Hvis l regnes i km, bliver σ_k spredningen på et nivellement over 1 km og kaldes kilometerspredningen. Almindeligt forekommende værdier af σ_k er:

- Teknisk linjenivellement: $\sigma_k \sim 5$ mm/$\sqrt{\text{km}}$
- Præcisionsnivellement: $\sigma_k \sim 0,5 - 1$ mm/$\sqrt{\text{km}}$

Bemærk: Ovenstående angår enkeltnivellement!

4.2 Vægtet gennemsnit

Antag målingerne $x_1, \ldots, x_n$ er udfald af uafhængige stokastiske variable $X_1, \ldots, X_n$ med fælles middelværdi μ men *forskellige* varianser $\sigma_i^2 = \mathbb{V}\mathrm{ar}(X_i)$, $i = 1, \ldots, n$.

Som estimat for μ anvendes det såkaldte *vægtede gennemsnit* $\bar{x}^*$, der er defineret ved

$$\bar{x}^* = \frac{p_1}{\sum_i p_i} x_1 + \cdots + \frac{p_n}{\sum_i p_i} x_n.$$

Her er $p_1, \ldots, p_n$ positive tal (vægte), der er valgt således, at vægtrelationen

$$p_1\sigma_1^2 = \ldots = p_n\sigma_n^2 \tag{4.6}$$

er opfyldt. Vægtrelationen medfører, at hvis eksempelvis σ_1^2 er dobbelt så stor som σ_2^2, så er vægten p_1 halvt så stor som p_2. Med andre ord er vægtene omvendt proportionale med de tilsvarende varianser,

$$p_i = \frac{k}{\sigma_i^2} \quad i = 1, \ldots, n,$$

for en positiv konstant k. Altså tillægges der, på naturlig vis, størst vægt til de observationer, der har mindst varians.

Sætning 19
Med henblik på at vurdere, hvor godt et estimat $\bar{x}^*$ er for μ, betragtes den tilsvarende stokastiske variabel $\bar{X}^*$, der er defineret ved udtrykket:

$$\bar{X}^* = \frac{p_1}{\sum_i p_i} X_1 + \ldots + \frac{p_n}{\sum_i p_i} X_n,$$

hvor $X_1, \ldots, X_n$ er defineret som i starten af afsnittet. Der gælder da følgende resultat:

1. *$\mathbb{E}[\bar{X}^*] = \mu$ og hermed er $\bar{X}^*$ en central estimator for μ.*
2. *Antag $p_1, \ldots, p_n$ er positive tal. Da er variansen for $\bar{X}^*$*

$$\mathbb{V}\mathrm{ar}\left[\frac{p_1}{\sum_i p_i} X_1 + \cdots + \frac{p_n}{\sum_i p_i} X_n\right]$$

mindst, når $p_1, \ldots, p_n$ opfylder vægtrelationen, dvs. når $p_1\sigma_1^2 = \cdots = p_n\sigma_n^2$.

Egenskaberne 1 og 2 i sætning 19 sikrer, at $\bar{x}^*$ er det bedst mulige centrale estimat for μ.

Vi begrænser os til at bevise sætning 19 i tilfældet, hvor $n = 2$. Vi kan fx antage, at vi har målt en vinkel μ med to teodolitter af forskellig fabrikat. Med den første teodolit har vi målt vinklen til x_1 gon, og med den anden teodolit er vinklen målt til x_2 gon. Vi får hermed følgende model

$$\begin{array}{cccc} X_1 & X_2 & \bar{X}^* = \frac{p_1}{p_1+p_2}X_1 + \frac{p_2}{p_1+p_2}X_2 \\ \downarrow & \downarrow & \downarrow \\ x_1 & x_2 & \bar{x}^* = \frac{p_1}{p_1+p_2}x_1 + \frac{p_2}{p_1+p_2}x_2 \end{array}$$

Her er p_1 og p_2 to positive tal, der opfylder vægtrelationen.

***Bevis** (Sætning 19)* Bevis for resultat 1 i sætning 19:

Det følger af sætning 3, at

$$\mathbb{E}[\bar{X}^*] = \mathbb{E}\left[\frac{p_1}{p_1+p_2}X_1 + \frac{p_2}{p_1+p_2}X_2\right] = \frac{p_1}{p_1+p_2}\mu + \frac{p_2}{p_1+p_2}\mu = \mu.$$

Dvs. $\bar{X}^*$ er en central estimator for middelværdien μ. Hermed er resultat 1 i sætning 19 bevist.

Bevis for resultat 2 i sætning 19: Betragt den stokastiske variabel $a_1X_1 + a_2X_2$ hvor a_1 og a_2 er to positive tal, således at $a_1 + a_2 = 1$. Vi vil nu bestemme a_1 (og dermed $a_2 = 1 - a_1$) således, at $\mathbb{V}\mathrm{ar}(a_1X_1 + a_2X_2)$ bliver mindst mulig. Da X_1 og X_2 antages at være uafhængige har vi jf. sætning 5, at variansen er givet ved

$$\mathbb{V}\mathrm{ar}(a_1X_1 + a_2X_2) = a_1^2\sigma_1^2 + (1-a_1)^2\sigma_2^2,$$

hvor $\sigma_1^2 = \mathbb{V}\mathrm{ar}[X_1]$ og $\sigma_2^2 = \mathbb{V}\mathrm{ar}[X_2]$.

Ved differentiation af dette udtryk mht. a_1 indses, at $\mathbb{V}\mathrm{ar}(a_1X_1 + a_2X_2)$ er mindst når

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

og hermed

$$a_2 = 1 - a_1 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

På den anden side har vi, at

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad \text{og} \quad a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2},$$

hvis og kun hvis

$$a_1\sigma_1^2 = a_2\sigma_2^2 \quad \text{og} \quad a_1 + a_2 = 1 \quad \text{(hvorfor?)}$$

Hermed opnår vi, at $\mathbb{V}\mathrm{ar}(a_1X_1 + a_2X_2)$ er mindst mulig, når $a_1\sigma_1^2 = a_2\sigma_2^2$ (og $a_1 + a_2 = 1$).

Indsættes $a_1 = \frac{p_1}{p_1+p_2}$ og $a_2 = \frac{p_2}{p_1+p_2}$ fås:

$$\mathbb{V}\mathrm{ar}\left(\frac{p_1}{p_1+p_2}X_1 + \frac{p_2}{p_1+p_2}X_2\right) \qquad \text{mindst for} \qquad p_1\sigma_1^2 = p_2\sigma_2^2$$

Det generelle tilfælde vises på tilsvarende vis. □

Bemærkning:

Antag vægtene $p_1, \ldots, p_n$ opfylder vægtrelationen $p_1\sigma_1^2 = \ldots = p_n\sigma_n^2$. Tallet σ_0 defineret ved

$$\sigma_0^2 = p_1\sigma_1^2 = \ldots = p_n\sigma_n^2$$

betegnes spredningen på en observation med vægt 1 eller spredningen på vægtenheden.

Som estimat for spredningen på vægtenheden σ_0 anvendes

$$s_0 = \sqrt{\frac{\sum_i p_i(x_i - \bar{x}^*)^2}{n-1}}.$$

Hvis alle observationer har vægten 1, svarer det vægtede gennemsnit til det sædvanlige gennemsnit, dvs. $\bar{x}^* = \bar{x}$. I dette tilfælde fås

$$s_0 = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}},$$

hvilket er det sædvanlige estimat for spredningen.

Variansen for det vægtede gennemsnit er

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\bar{X}^*] &= \mathbb{V}\mathrm{ar}\left[\frac{p_1}{\sum_i p_i}X_i + \cdots + \frac{p_n}{\sum_i p_i}X_n\right] \\
&= \frac{1}{(\sum_i p_i)^2}\sum_i p_i^2\sigma_i^2 = \frac{\sum_i p_i\sigma_0^2}{(\sum_i p_i)^2} \\
&= \frac{\sigma_0^2}{\sum_i p_i}
\end{aligned}$$

Dermed er et estimat for variansen på det vægtede gennemsnit

$$\frac{s_0^2}{\sum_i p_i},$$

hvilket medfører, at $\frac{s_0}{\sqrt{\sum_i p_i}}$ bliver et estimat for spredningen på det vægtede gennemsnit.

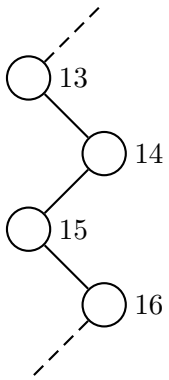
Idet

$$\mathbb{V}\mathrm{ar}[X_i] = \sigma_i^2 = \frac{\sigma_0^2}{p_i}$$

er et estimat for variansen for den i'te måling givet ved

$$\frac{s_0^2}{p_i},$$

og dermed bliver $s_0/\sqrt{p_i}$ et estimat for spredningen på den i'te måling.



Figur 4.2: Illustration til eksempel 25.

Eksempel 25
Antag, at vi måler polygonvinklerne i 13-16 i figur 4.2 med samme teodolit under samme omstændigheder. Det betyder, at vinklerne målt med én sats har samme spredning σ_v.

Måler vi et forskelligt antal satser i punkterne, vil middelsatsens spredning $\sigma_{m,i}$ variere som funktion af satsantallet. Lad os antage, at vi måler med de i skemaet angivne antal satser:

Punkt	Antal satser
13	2
14	3
15	4
16	6

Tabel 4.1: Eksempel: Antal satser

Da fås følgende varianser for middelsatserne:

$$\sigma_{m,13}^2 = \frac{\sigma_v^2}{2}; \qquad \sigma_{m,14}^2 = \frac{\sigma_v^2}{3}$$
$$\sigma_{m,15}^2 = \frac{\sigma_v^2}{4}; \qquad \sigma_{m,16}^2 = \frac{\sigma_v^2}{6}$$

Skal middelsatsernes værdier indgå i en udjævning, må de tildeles vægte efter vægtrelationen:

$$p_1\sigma_{m,13}^2 = p_2\sigma_{m,14}^2 = p_3\sigma_{m,15}^2 = p_4\sigma_{m,16}^2,$$

eller

$$p_1\frac{\sigma_v^2}{2} = p_2\frac{\sigma_v^2}{3} = p_3\frac{\sigma_v^2}{4} = p_4\frac{\sigma_v^2}{6}.$$

Vælges vægtene $p_1 = 2$, $p_2 = 3$, $p_3 = 4$ og $p_4 = 6$, fås, at variansen på vægtenheden svarer til variansen på en måling af én sats, dvs. én sats er vægtenheden.

Vægtene kan vælges på uendelige mange måder, blot vægtrelationen er opfyldt. Andre mulige vægte er

Punkt	13	14	15	16	Vægtenhed
	2	3	4	6	1 sats
Vægte	1	$\frac{3}{2}$	2	3	2 satser
	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	1	6 satser

◊

Øvelse 8
I et net måles med sekundteodolit, hvis retningsspredning kan antages at være $\sigma_{r,1} = 0,0008$ *gon. Der måles overalt med to satser. Målingerne ønskes suppleret, og der er kun en mindre teodolit til rådighed. Teodolitten antages at have en retningsspredning* $\sigma_{r,2} = 0,0020$ *gon. Der ønskes målt så mange satser, at middelsatserne fra suppleringsmålingerne kan indgå med samme vægt som de oprindelige. Hvor mange satser skal der måles med?*

4.3 Fordeling af slutfejl

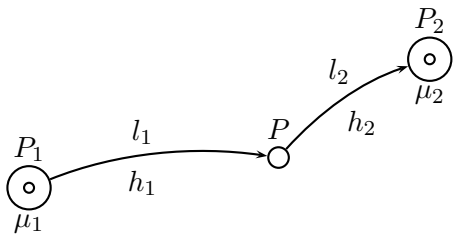
Fordeling af slutfejl forekommer hyppigt i landmålingen. Fra beregning af polygoner er det velkendt, at vinkelfejl normalt fordeles ligeligt på alle vinkler. Vi vil i dette afsnit begrunde hvorfor. Men først vil vi undersøge et tilsvarende simpelt problem i nivellement.

4.3.1 Fordeling af slutfejl i forbindelse med nivellement

Antag, at der er udført et nivellement mellem to punkter P_1 og P_2 med kendte koter μ_1 og μ_2, se figur 4.3. Vi vil gennem nivellement beregne koten μ til det vilkårlige punkt P, der har den målte højdeforskel h_1 til P_1 og h_2 til P_2. Længden fra P_1 til P er l_1, og fra P_2 til P er længden l_2. Vi antager, at h_1 og h_2 er udfald af uafhængige stokastiske variable H_1 og H_2 med middelværdier $\mu - \mu_1$ og $\mu_2 - \mu$ og varianser $l_1\sigma_k^2$ og $l_2\sigma_k^2$, jf. (4.5). Vi lader videre $X_1 = \mu_1 + H_1$ og $X_2 = \mu_2 - H_2$. Det følger da, at X_1 og X_2 også er uafhængige med varianser $l_1\sigma_k^2$ og $l_2\sigma_k^2$ og fælles middelværdi μ.

Givet observationer $x_1 = \mu_1 + h_1$ og $x_2 = \mu_2 - h_2$ kan vi da anvende det vægtede gennemsnit

$$\bar{x}^* = \frac{p_1}{p_1 + p_2}x_1 + \frac{p_2}{p_1 + p_2}x_2$$



Figur 4.3: Nivellement mellem punkterne P_1 og P_2 via punktet P.

som estimat for μ. Her er p_1 og p_2 vægte, der opfylder vægtrelationen, dvs.

$$p_1 \mathbb{V}\mathrm{ar}(X_1) = p_2 \mathbb{V}\mathrm{ar}(X_2)$$

eller (se ovenfor)

$$p_1 l_1 \sigma_k^2 = p_2 l_2 \sigma_k^2. \tag{4.7}$$

Et valg af vægte, der opfylder (4.7), er

$$p_1 = \frac{1}{l_1} \quad \text{og} \quad p_2 = \frac{1}{l_2}.$$

Med dette valg bliver det vægtede gennemsnit

$$\bar{x}^* = \frac{\frac{1}{l_1}}{\frac{1}{l_1} + \frac{1}{l_2}} x_1 + \frac{\frac{1}{l_2}}{\frac{1}{l_1} + \frac{1}{l_2}} x_2. \tag{4.8}$$

På grund af uundgåelig målefejl, er $h_1 + h_2$ ikke lig $\mu_2 - \mu_1$. Defineres slutfejlen r som forskellen

$$r = \mu_2 - \mu_1 - (h_1 + h_2)$$

fås

$$\mu_2 - h_2 = \mu_1 + h_1 + r \tag{4.9}$$

eller

$$x_2 = x_1 + r$$

Erstatter vi x_2 med $x_1 + r$ i det vægtede gennemsnit (4.8), fås

$$\bar{x}^* = x_1 + \frac{\frac{1}{l_2}}{\frac{1}{l_1} + \frac{1}{l_2}} r,$$

og dermed

$$\bar{x}^* = x_1 + \frac{l_1}{l_1 + l_2}.$$

Dvs. at vi har omskrevet det vægtede gennemsnit til en sum af x_1 og en andel af slutfejlen, der svarer til vejlængden l_1s andel af den totale vejlængde. På tilsvarende vis fås

$$\bar{x}^* = x_2 - \frac{l_2}{l_1 + l_2} r.$$

Ovenstående kan formuleres som at slutfejlen fordeles på henholdsvis x_1 og x_2 proportionalt med vejlængden.

4.3.2 Fordeling af slutfejl for vinkelmålinger

Vi vender nu tilbage til måling af vinkler i polygoner. Betragt trekanten i figur 1.1 med vinklerne α, β og γ (sande størrelser). Vinkel α er målt til x_α gon, vinkel β er målt til x_β gon, og vinkel γ er målt til x_γ gon. Vi har målt alle tre vinkler med samme antal satser

Model:

$$\begin{array}{ccc} X_\alpha & X_\beta & X_\gamma \\ \downarrow & \downarrow & \downarrow \\ x_\alpha & x_\beta & x_\gamma \end{array}$$

Mere specifikt antager vi, at X_α, X_β og X_γ er tre uafhængige stokastiske variable, der har antaget værdierne x_α, x_β og x_γ henholdsvis. Middelværdierne for X_α, X_β og X_γ er

$$\mathbb{E}(X_\alpha) = \alpha, \;\; \mathbb{E}(X_\beta) = \beta, \text{ og } \mathbb{E}(X_\gamma) = \gamma$$

og varianserne er $\mathbb{V}\mathrm{ar}(X_\alpha) = \mathbb{V}\mathrm{ar}(X_\beta) = \mathbb{V}\mathrm{ar}(X_\gamma) = \sigma^2$, dvs. alle vinkler er målt med lige stor præcision.

Lad nu $Y_\alpha = 200 - X_\beta - X_\gamma$. Dermed er X_α og Y_α to uafhængige stokastiske variable, der har antaget værdierne x_α og $y_\alpha = 200 - x_\beta - x_\gamma$, henholdsvis. Middelværdierne for X_α og Y_α er

$$\mathbb{E}[X_\alpha] = \alpha \quad \text{og} \quad \mathbb{E}[Y_\alpha] = \mathbb{E}[200 - X_\beta - X_\gamma] = 200 - \beta - \gamma = \alpha,$$

hvor den sidste lighed er en konsekvens af, at $\alpha + \beta + \gamma = 200$. Varianserne er

$$\mathbb{V}\mathrm{ar}(X_\alpha) = \sigma^2 \text{ og } \mathbb{V}\mathrm{ar}(Y_\alpha) = \mathbb{V}\mathrm{ar}(200 - X_\beta - X_\gamma) = \mathbb{V}\mathrm{ar}(X_\beta) + \mathbb{V}\mathrm{ar}(X_\gamma) = 2\sigma^2.$$

Da de to bestemmelser (målinger) af α, dvs. x_α og y_α, ikke er tilknyttet samme varians, estimeres α ved et vægtet gennemsnit til

$$\bar{x}^* = \frac{p_1}{p_1 + p_2} x_\alpha + \frac{p_2}{p_1 + p_2} y_\alpha,$$

hvor p_1 og p_2 er vægte, der opfylder vægtrelationen

$$p_1 \mathbb{V}\mathrm{ar}(X_\alpha) = p_2 \mathbb{V}\mathrm{ar}(Y_\alpha).$$

Idet $\mathbb{V}\mathrm{ar}(X_\alpha) = \sigma^2$ og $\mathbb{V}\mathrm{ar}(Y_\alpha) = 2\sigma^2$, kan vægtrelationen omskrives til

$$p_1\sigma^2 = p_2 2\sigma^2.$$

Som vægte kan vi hermed anvende $p_1 = 2$ og $p_2 = 1$. Dvs.

$$\bar{x}^* = \frac{2}{3}x_\alpha + \frac{1}{3}y_\alpha. \tag{4.10}$$

På grund af tilfældige fejl gælder der ikke $x_\alpha + x_\beta + x_\gamma = 200$. Vi definerer følgelig korrektionen r_v (slutfejlen) som

$$r_v = 200 - (x_\alpha + x_\beta + x_\gamma),$$

som kan omskrives til

$$200 - x_\beta - x_\gamma = x_\alpha + r_v.$$

Erstatter vi $y_\alpha = 200 - x_\beta - x_\gamma$ med $x_\alpha + r_v$ i (4.10) bliver det vægtede gennemsnit

$$\bar{x}^* = x_\alpha + \frac{1}{3}r_v.$$

Tilsvarende fås estimaterne $x_\beta + \frac{1}{3}r_v$ og $x_\gamma + \frac{1}{3}r_v$ for henholdsvis β og γ. Slutfejlen r_v skal altså fordeles ligeligt på alle vinkler uanset størrelse.

4.4 Dobbeltmålinger

I landmålingen forekommer ofte, at man danner et antal gennemsnit af par af målinger. Således måles polygonsider tit dobbelt, og nivellement udføres som frem- og tilbagenivellement. Udgangspunktet for dette afsnit er altså følgende:

$$\begin{gathered} x_{11}, x_{12} \text{ er to målinger af størrelsen } \mu_1 \\ x_{21}, x_{22} \text{ er to målinger af størrelsen } \mu_2 \\ \vdots \\ x_{n1}, x_{n2} \text{ er to målinger af størrelsen } \mu_n \end{gathered}$$

Alle målinger er udført med samme målekvalitet. Målingerne er udført uafhængigt af hinanden.

$$\begin{gathered} \text{Som estimat for } \mu_1 \text{ anvendes } \bar{x}_1 = \tfrac{1}{2}(x_{11} + x_{12}) \\ \text{Som estimat for } \mu_2 \text{ anvendes } \bar{x}_2 = \tfrac{1}{2}(x_{21} + x_{22}) \\ \vdots \\ \text{Som estimat for } \mu_n \text{ anvendes } \bar{x}_n = \tfrac{1}{2}(x_{n1} + x_{n2}) \end{gathered}$$

Vi formulerer en statistisk model:

$$\begin{array}{cccc} X_{11}, X_{12} & X_{21}, X_{22} & \dots & X_{n1}, X_{n2} \\ \downarrow & \downarrow & & \downarrow \\ x_{11}, x_{12} & x_{21}, x_{22} & \dots & x_{n1}, x_{n2} \end{array}$$

Mere præcist antager vi, at $X_{11}, X_{12}, X_{21}, X_{22}, \dots, X_{n1}, X_{n2}$ er $2n$ uafhængige stokastiske variable, der har antaget værdierne $x_{11}, x_{12}, x_{21}, x_{22}, \dots, x_{n1}, x_{n2}$. Middelværdierne er

$$\mathbb{E}(X_{11}) = \mathbb{E}(X_{12}) = \mu_1,\ \mathbb{E}(X_{21}) = \mathbb{E}(X_{22}) = \mu_2,\ \dots,\ \mathbb{E}(X_{n1}) = \mathbb{E}(X_{n2}) = \mu_n,$$

og varianserne er ens

$$\mathbb{V}\mathrm{ar}(X_{11}) = \mathbb{V}\mathrm{ar}(X_{12}) = \mathbb{V}\mathrm{ar}(X_{21}) = \mathbb{V}\mathrm{ar}(X_{22}) = \cdots = \mathbb{V}\mathrm{ar}(X_{n1}) = \mathbb{V}\mathrm{ar}(X_{n2}) = \sigma^2.$$

De ovenfor nævnte gennemsnit $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$ er hermed antaget af de stokastiske variable

$$\bar{X}_1 = \frac{1}{2}(X_{11} + X_{12}), \bar{X}_2 = \frac{1}{2}(X_{21} + X_{22}), \dots, \bar{X}_n = \frac{1}{2}(X_{n1} + X_{n2}).$$

Middelværdien for $\bar{X}_1$ er

$$\mathbb{E}[\bar{X}_1] = E\left[\frac{1}{2}(X_{11} + X_{12})\right] = \frac{1}{2}(\mu_1 + \mu_1) = \mu_1.$$

Da denne udregning gælder for alle n gennemsnit, har vi

$$\mathbb{E}[\bar{X}_1] = \mu_1, \mathbb{E}[\bar{X}_2] = \mu_2, \dots, \mathbb{E}[\bar{X}_n] = \mu_n.$$

Variansen for $\bar{X}_1$ er

$$\mathbb{V}\mathrm{ar}[\bar{X}_1] = \mathbb{V}\mathrm{ar}\left[\frac{1}{2}(X_{11} + X_{12})\right] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}.$$

Da denne udregning gælder for alle n gennemsnit har vi

$$\mathbb{V}\mathrm{ar}[\bar{X}_1] = \frac{\sigma^2}{2}, \mathbb{V}\mathrm{ar}[\bar{X}_2] = \frac{\sigma^2}{2}, \dots, \mathbb{V}\mathrm{ar}[\bar{X}_n] = \frac{\sigma^2}{2}.$$

Idet σ^2 var variansen på den enkelte måling, bliver $\frac{\sigma^2}{2}$ lig variansen på gennemsnittet af hvert målepar.

Med henblik på at bestemme bedst mulige estimat for σ^2, betragtes følgende model:

$$\begin{array}{cccc} Y_1 = X_{11} - X_{12} & Y_2 = X_{21} - X_{22} & \dots & Y_n = X_{n1} - X_{n2} \\ \downarrow & \downarrow & & \downarrow \\ y_1 = x_{11} - x_{12} & y_2 = x_{21} - x_{22} & \dots & y_n = x_{n1} - x_{n2} \end{array}$$

Således er $Y_1, Y_2, \ldots, Y_n$ n uafhængige stokastiske variable, der har antaget værdierne $y_1, y_2, \ldots, y_n$. Mere interessant er, at modsat $\bar{X}_1 \ldots, \bar{X}_n$, så har $Y_1, Y_2, \ldots, Y_n$ alle kendt middelværdi! Middelværdien for Y_1 er

$$\mathbb{E}[Y_1] = \mathbb{E}\left[\frac{1}{2}(X_{11} - X_{12})\right] = \frac{1}{2}(\mu_1 - \mu_1) = 0.$$

Da denne udregning gælder for alle Y_i'er har vi

$$\mathbb{E}[Y_1] = 0, \ldots, \mathbb{E}[Y_n] = 0.$$

Variansen for Y_1 er

$$\mathbb{V}\mathrm{ar}[Y_1] = \mathbb{V}\mathrm{ar}[X_{11} - X_{12}] = \sigma^2 + \sigma^2 = 2\sigma^2.$$

Da denne udregning gælder for alle Y_i'er har vi

$$\mathbb{V}\mathrm{ar}[Y_1] = 2\sigma^2, \mathbb{V}\mathrm{ar}[Y_2] = 2\sigma^2, \ldots, \mathbb{V}\mathrm{ar}[Y_n] = 2\sigma^2.$$

Vi har altså n tal $y_1, \ldots, y_n$, der er værdier antaget af n uafhængige stokastiske variable $Y_1, \ldots, Y_n$, der alle har kendt middelværdi 0 og ukendt varians $2\sigma^2$.

Da middelværdien er kendt gælder jf. sætning 18 (side 39), at det optimale estimat for $2\sigma^2$ *ikke* er det sædvanlige variansestimat

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

men derimod

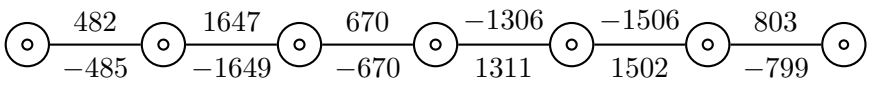
$$2\hat{s}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - 0)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - x_{i2})^2.$$

Estimatet for variansen, σ^2, på den enkelte måling er altså

$$\hat{s}_Y^2 = \frac{1}{2n}\sum_{i=1}^{n}(x_{i1} - x_{i2})^2.$$

Øvelse 9

Der foretages frem- og tilbagenivellement på 6 "lige lange"strækninger (ca. 120 m), se figur 4.4. De 12 enkeltnivellementer anses for lige gode. Spredningen på den enkelte måling betegnes σ. Bestem et estimat for σ.



Figur 4.4: Frem- og tilbagenivellementer over 6 strækninger.

Kapitel 5

Fejlforplantning

I landmålingen vil vi hyppigt stå over for den opgave at bestemme en størrelse, som ikke direkte kan måles. Vi må måle andre størrelser, hvoraf den søgte kan beregnes. Vil vi fx. bestemme arealet af en trekant, kan vi måle 2 sider og den mellemliggende vinkel. Arealet t kan herefter beregnes som:

$$t = \frac{1}{2}ab\sin C$$

Opfattes observationerne af de to sider og vinklen som stokastiske variable med kendte varianser, er vi interesseret i at bestemme, hvorledes disse varianser influerer på bestemmelsen af trekantens areal.

Generelt vil vi vha. *fejlforplantningslove* bestemme, hvorledes observationsfejl influerer på funktioner af de størrelser, der observeres. Vi vil derfor i det følgende finde tilnærmede udtryk for middelværdier og varianser for $Y_1, \ldots, Y_m$ givet ved reelle og differentiable funktioner af n stokastiske variable $X_1, X_2, \ldots, X_n$:

$$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_n) \\ &\vdots \\ Y_m &= g_m(X_1, \ldots, X_n). \end{aligned}$$

5.1 Uafhængige stokastiske variable

Er $X_1, X_2, \ldots, X_n$ indbyrdes uafhængige observationer, er alle kovarianser nul:

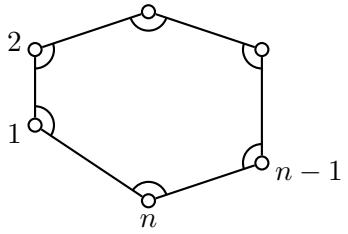
$$\mathrm{Cov}(X_i,\ X_j) = 0 \text{ for } 1 \leq i < j \leq n.$$

Lad Y være en linearkombination af de n uafhængige stokastiske variable:

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

Hvis de tilsvarende varianser er $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, får vi, jf. sætning 5,

$$\mathbb{V}\mathrm{ar}[Y] = a_1\sigma_1^2 + a_2\sigma_2^2 + \ldots + a_n\sigma_n^2. \tag{5.1}$$



Figur 5.1: Vinkler i polygon med n vinkler.

Eksempel 26
Betragt den stokastiske variabel Y givet ved

$$Y = X_1 + X_2,$$

hvor X_1 og X_2 er uafhængige stokastiske variable. Variansen for Y er $\sigma_Y^2 = \sigma_1^2 + \sigma_2^2$, hvor $\sigma_i^2 = \mathbb{V}\mathrm{ar}[X_i]$. Hvis $\sigma_1 = \sigma_2 = \sigma$, er variansen for Y givet ved $\sigma_Y^2 = 2\sigma^2$, hvilket medfører, at standardafvigelsen for Y er givet ved $\sigma_Y = \sqrt{2}\sigma$.

Antag nu, at Y er summen af n uafhængige stokastiske variable, dvs.

$$Y = X_1 + X_2 + \ldots + X_n.$$

Da er variansen for Y

$$\sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2 = \sum_{i=1}^{n} \sigma_i^2.$$

Hvis observationerne er lige gode, dvs. $\sigma_i = \sigma$ for alle i, fås

$$\sigma_Y^2 = n\sigma^2 \quad \Rightarrow \quad \sigma_Y = \sqrt{n}\sigma.$$

Anvendelser

En vinkel beregnes altid som differensen mellem to observerede retninger

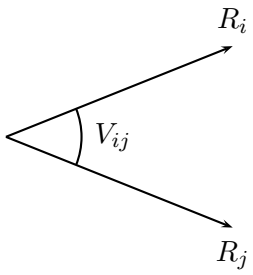
$$V_{ij} = R_i - R_j,$$

se figur 5.2. Hvis $\sigma_i = \sigma_j = \sigma_R$, opnår vi $\sigma_V = \sqrt{2}\sigma_R$. Spredningen på vinklen V_{ij} er $\sqrt{2}$ gange større end spredningen på hver af retningerne.

Spredningen på en vinkelsum i en lukket polygon med n vinkler er derfor $\sigma_s = \sqrt{2n} \cdot \sigma_R$, hvis alle retninger er målt med samme spredning σ_R.

Ved et geometrisk nivellement bestemmes højdeforskellen i den enkelte opstilling ved $h = t - f$, hvor t og f er stadieaflæsningerne ved henholdsvis tilbagesigte og fremsigte. Hvis stadieaflæsningen regnes lige gode med spredningen σ_a, er $\sigma_h^2 = 2\sigma_a^2$. Hvis der til bestemmelse af en højdeforskel H medgår lige gode højdeforskelle bestemt ved n opstillinger fås

$$\sigma_H^2 = 2n\sigma_a^2$$



Figur 5.2: To retninger, R_i og R_j, og den mellemliggende vinkel V_{ij}.

5.2 Linearisering

I afsnit 2.6 så vi, hvordan en funktion af én variabel kan lineariseres. Antag $X_1, \ldots, X_n$ er uafhængige med middelværdier $\mu_1, \ldots, \mu_n$ og varianser $\sigma_1^, \ldots, \sigma_n^2$. Hvis funktionen $Y = g(X_1, X_2, \ldots, X_n)$ ikke er lineær, men differentiabel, ved vi, at Y kan lineariseres. Lineariserer vi Y omkring punktet $(\mu_1, \ldots, \mu_n)$ fås et tilnærmet udtryk for Y af formen

$$Y \approx g(\mu_1, \mu_2, \ldots, \mu_n) + \frac{\partial g}{\partial X_1}(X_1 - \mu_1) + \frac{\partial g}{\partial X_2}(X_2 - \mu_2) + \cdots + \frac{\partial g}{\partial X_n}(X_n - \mu_n) \tag{5.2}$$

Notationen $\frac{\partial g}{\partial X_i}$ skal forstås som den i'te partielle afledede beregnet i punktet $(\mu_1, \ldots, \mu_n)$. I praksis kender vi ikke $\mu_1, \ldots, \mu_n$, hvorfor vi i stedet bruger estimater, typisk $\bar{x}_1, \ldots, \bar{x}_n$.

Et tilnærmet udtryk for $\mathbb{E}[Y]$ er

$$\begin{aligned}
\mathbb{E}[Y] &\approx \mathbb{E}\left[g(\mu_1, \mu_2, \ldots, \mu_n) + \frac{\partial g}{\partial X_1}(X_1 - \mu_1) + \cdots + \frac{\partial g}{\partial X_n}(X_n - \mu_n)\right] \\
&= g(\mu_1, \mu_2, \ldots, \mu_n) + \frac{\partial g}{\partial X_1}\mathbb{E}(X_1 - \mu_1) + \cdots + \frac{\partial g}{\partial X_n}\mathbb{E}(X_n - \mu_n) \\
&= g(\mu_1, \mu_2, \ldots, \mu_n)
\end{aligned}$$

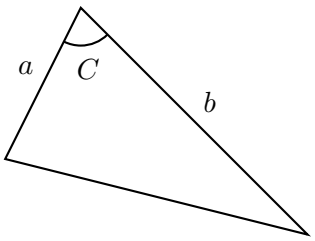
Med hensyn til variansen fås

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y] &\approx \left(\frac{\partial g}{\partial X_1}\right)^2 \mathbb{V}\mathrm{ar}(X_1) + \left(\frac{\partial g}{\partial X_2}\right)^2 \mathbb{V}\mathrm{ar}(X_2) + \ldots + \left(\frac{\partial g}{\partial X_n}\right)^2 \mathbb{V}\mathrm{ar}(X_n) \\
&= \left(\frac{\partial g}{\partial X_1}\right)^2 \sigma_1^2 + \left(\frac{\partial g}{\partial X_2}\right)^2 \sigma_2^2 + \ldots + \left(\frac{\partial g}{\partial X_n}\right)^2 \sigma_n^2.
\end{aligned} \tag{5.3}$$

Bemærk at (5.3) også kan anvendes for lineære funktioner, dvs. har generel gyldighed for funktioner af indbyrdes uafhængige stokastiske variable.

Eksempel 27

For trekanten i figur 5.3 måles siderne a og b samt vinklen C. Arealet er $t =$



Figur 5.3: Endnu en trekant.

$\frac{1}{2}ab\sin C$. Observationerne, der opfattes som indbyrdes uafhængige, er målt til

$$\begin{aligned} a &= 115,53 \text{ m} \\ b &= 152,17 \text{ m} \\ C &= 93,273 \text{ gon} \end{aligned}$$

Observationerne er associerede med følgende usikkerheder

$$\begin{aligned} \sigma_a = \sigma_b &= 1 \text{ cm} \\ \sigma_c &= 0,002 \text{ gon} \end{aligned}$$

Estimatet $\hat{t}$ af arealet opnås ved at indsætte de observerede værdier i formlen for arealet:

$$t \approx \hat{t} = \frac{1}{2}115,53 \cdot 152,17 \cdot \sin(93,273) = 8741 \text{ m}^2$$

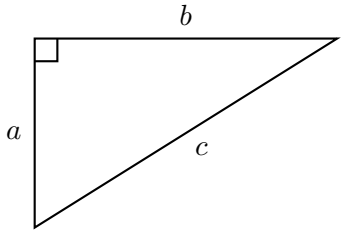
Udtrykkene for de partielle afledede af t samt de konkrete udregnede værdier af de partielle afledede er:

$$\begin{aligned} \frac{\partial t}{\partial a} &= \frac{1}{2}b\sin C = \frac{t}{a} & \Rightarrow \frac{\partial t}{\partial a} &= 75,66 \text{ m} \\ \frac{\partial t}{\partial b} &= \frac{1}{2}a\sin C = \frac{t}{b} & \Rightarrow \frac{\partial t}{\partial b} &= 57,44 \text{ m} \\ \frac{\partial t}{\partial C} &= \frac{1}{2}ab\cos C = \frac{t}{\tan C} & \Rightarrow \frac{\partial t}{\partial C} &= 927,1 \text{ m}^2 \end{aligned}$$

Indsættes i (5.3) får vi et tilnærmet udtryk for variansen for estimatet $\hat{t}$ af t:

$$\begin{aligned} \sigma_{\hat{t}}^2 &\approx (75,66)^2 \cdot (0,01)^2 \text{ m}^4 + (57,44)^2 \cdot (0,01)^2 \text{ m}^4 + (927,1)^2 \left(0,002\frac{\pi}{200}\right)^2 \text{ m}^4 \\ &= 0,57 \text{ m}^4 + 0,33 \text{ m}^4 + 0,0008 \text{ m}^4 \\ \sigma_{\hat{t}}^2 &\approx 0,90 \text{ m}^4 \Rightarrow \sigma_T \approx 0,95 \text{ m}^2. \end{aligned}$$

Hvor kommer faktoren $\pi/200$ fra? De trigonometriske funktioner er defineret for vinkler målt i radianer. Når vi normalt anvender de trigonometriske



Figur 5.4: En retvinklet trekant

funktioner på vinkler målt i gon, går det godt, fordi lommeregneren "usynligt" konverterer til radianer. Skal vi derimod differentiere trigonometriske funktioner, går det galt, hvis vi anvender vinkler målt i andet end radianer. Antag α er en vinkel målt i gon. Normalt ville vi skrive $\sin(\alpha)$, men det korrekte ville være at skrive

$$\sin\left(\alpha\frac{\pi}{200}\right). \tag{5.4}$$

Differentierer vi nu (5.4) mht. α får vi

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\sin\left(\frac{\alpha\pi}{200}\right) = \cos\left(\frac{\alpha\pi}{200}\right)\frac{\pi}{200},$$

hvilket forklarer faktoren $\pi/200$ i udregningerne ovenfor.

Hvis vi tror på, at målingerne af a, b og C er normalfordelte, er et approksimativt 95% konfidensinterval for t givet ved

$$\hat{t} \pm 1,96\hat{\sigma}_{\hat{t}} = 8741 \pm 1,86.$$

Øvelse 10
I en retvinklet trekant måles a og b uafhængigt med samme spredning σ, se figur 5.4.

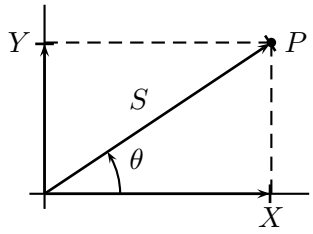
1. *Find spredningen på $c = \sqrt{a^2 + b^2}$.*

Arealet er $t = \frac{1}{2}ab$

2 *Find værdien $\sigma_{\hat{t}}$ med værdierne af a og b fra eksempel 27.*
Sammenlign og kommenter forskellen mellem den udregnede spredning og spredningen i eksempel 27.

Øvelse 11
I det følgende betragter vi to variable X og Y:

$$Y = S\sin\theta \quad og$$
$$X = S\cos\theta,$$



Figur 5.5: Illustration til øvelse 11.

hvor S og θ er indbyrdes uafhængige observationer, se figur 5.5. Længden S er målt med edm til værdien 215,64 m, og $\sigma_S = 1$ cm. Vinklen θ er målt med teodolit til værdien 62,263 gon, $\sigma_\theta = 0,002$ gon.

1. *Beregn estimater for Y og X.*

2. *Bestem ved hjælp af* (5.3) *den tilnærmede spredning på Y og X.*

5.3 Den generelle fejlforplantningslov

Indtil nu har vi fundet (tilnærmede) udtryk for middelværdi og varians for en eller flere funktioner af uafhængige stokastiske variable. I øvelse 11 fandt vi således tilnærmede middelværdier og varianser for koordinaterne for punktet P, når vi har målt vinkel og afstand. En nærmere analyse viser, at koordinaterne X og Y er *afhængige* variable. Vi vil i dette afsnit finde et tilnærmet udtryk for kovariansen mellem X og Y. Antag, at vi vil finde afstanden mellem punktet P og et andet punkt. Denne afstand er en funktion af de afhængige stokastiske variable X og Y. Da antagelsen om uafhængighed ikke er opfyldt, kan vi ikke bruge (5.3) til finde et tilnærmet udtryk for variansen. I dette og andre tilfælde, hvor de variable, der indgår i funktionerne, er korrelerede, vil kovariansen få indflydelse på fejlforplantningen.

Vi starter med at betragte et eksempel, hvor to variable, Y_1 og Y_2, begge er lineære funktioner af de samme to *uafhængige* stokastiske variable X_1 og X_2. Selvom X_1 og X_2 er uafhængige er kovariansen mellem Y_1 og Y_2 typisk ikke nul. I det følgende eksempel finder vi kovariansen mellem Y_1 og Y_2 i et konkret tilfælde.

Eksempel 28
Lad X_1 og X_2 være uafhængige stokastiske variable med varianserne σ_1^2 og σ_2^2. Med udgangspunkt i X_1 og X_2 defineres Y_1 og Y_2 som

$$Y_1 = 2X_1 - 3X_2 \quad \text{og} \tag{5.5}$$
$$Y_2 = 3X_1 + 2X_2. \tag{5.6}$$

Vi vil undersøge, om Y_1 og Y_2 er korrelerede eller ej. Ifølge sætning 12 har

vi

$$\begin{aligned}\mathbb{V}\text{ar}[Y_1] &= 4\sigma_1^2 + 9\sigma_2^2 \\ \mathbb{V}\text{ar}[Y_2] &= 9\sigma_1^2 + 4\sigma_2^2 \\ \mathbb{V}\text{ar}[Y_1 + Y_2] &= \mathbb{V}\text{ar}(5X_1 - X_2) = 25\sigma_1^2 + \sigma_2^2\end{aligned}$$

Sætning 11 kan omskrives til

$$\mathbb{V}\text{ar}(Y_1 + Y_2) = \mathbb{V}\text{ar}(Y_1) + \mathbb{V}\text{ar}(Y_2) + 2\text{Cov}(Y_1, Y_2).$$

Ved indsættelse får vi

$$25\sigma_1^2 + \sigma_2^2 = 4\sigma_1^2 + 9\sigma_2^2 + 9\sigma_1^2 + 4\sigma_2^2 + 2\text{Cov}(Y_1, Y_2).$$

Hvis vi isolerer kovariansen, får vi

$$\text{Cov}(Y_1, Y_2) = 6(\sigma_1^2 - \sigma_2^2).$$

Kovariansen er således forskellig fra nul, hvis $\sigma_1^2 \neq \sigma_2^2$, dvs. Y_1 og Y_2 er korrelerede, hvis $\sigma_1^2 \neq \sigma_2^2$ og ukorrelerede ellers. $\Diamond$

En alternativ tilgang er at anvende matrix-formuleringen i afsnit 2.9. Specifikt kan funktionerne (5.5) og (5.6) omskrives til

$$\mathbf{Y} = \mathbf{A}\mathbf{X},$$

hvor

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix} \quad \text{og} \quad \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Jf. sætning 14 er variansen for $\mathbf{Y}$ givet ved

$$\begin{aligned}\mathbb{V}\text{ar}[\mathbf{Y}] &= \mathbf{A}\mathbf{K}_\mathbf{X}\mathbf{A}^T \\ &= \begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ -3 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 4\sigma_1^2 + 9\sigma_2^2 & 6(\sigma_1^2 - \sigma_2^2) \\ 6(\sigma_1^2 - \sigma_2^2) & 9\sigma_1^2 + 4\sigma_2^2 \end{bmatrix} \end{aligned} \tag{5.7}$$

I diagonalen i kovariansmatricen (5.7) kan vi aflæse varianserne

$$\mathbb{V}\text{ar}[Y_1] = 4\sigma_1^2 + 9\sigma_2^2 \quad \text{og} \quad \mathbb{V}\text{ar}[Y_2] = 9\sigma_1^2 + 4\sigma_2^2,$$

hvilket stemmer overnes med, hvad vi fandt i eksempel 28 ovenfor. Kovariansen mellem Y_1 og Y_2 aflæses til at være $\text{Cov}(Y_1, Y_2) = 6(\sigma_1^2 - \sigma_2^2)$, hvilket også stemmer overens med resultaterne i eksempel 28. I situationer, der involverer lidt flere variable end i eksempel 28, er matrix-tilgangen generelt mere bekvem. I praksis håndteres matrixberegninger let vha. f.eks. matlab.

Vi betragter nu situationen, hvor et enkelt Y er en funktion af indbyrdes afhængige stokastiske variable $X_1, \ldots, X_n$. Er Y en lineær funktion af $X_1, \ldots, X_n$,

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

har vi fra sætning 13, at variansen for Y er

$$\mathbb{V}\text{ar}\left[a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n\right] = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i<j} a_ia_j\sigma_{ij}, \tag{5.8}$$

hvor σ_{ij} er kovariansen mellem X_i og X_j.

Er funktionen $Y = g(X_1, \ldots, X_n)$ ikke lineær, men differentiabel, kan vi stadig finde et tilnærmet udtryk for variansen. Specifikt finder vi et tilnærmet udtryk for variansen for Y ved at finde variansen for den generelle linearisering i (5.2) vha. (5.8), hvor

$$a_1 = \frac{\partial g}{\partial X_1}, \ldots, a_n = \frac{\partial g}{\partial X_n},$$

dvs. et tilnærmet udtryk for variansen for Y er

$$\mathbb{V}\text{ar}[Y] \approx \sum_{i=1}^{n} \left(\frac{\partial g}{\partial X_i}\right)^2 \sigma_i^2 + 2\sum_{i<j} \left(\frac{\partial g}{\partial X_i}\right)\left(\frac{\partial g}{\partial X_j}\right)\sigma_{ij}. \tag{5.9}$$

I det følgende eksempel finder vi variansen for en stokastisk variabel Y, der er en ikke-lineær funktion af to afhængige variable.

Eksempel 29
Lad X_1 og X_2 være stokastiske variable med middelværdier μ_1 og μ_2, og varianser σ_1^2 og σ_2^2. Kovariansen mellem X_1 og X_2 er forskellig fra nul, dvs. $\text{Cov}(X_1,\ X_2) \neq 0$. Nedenfor benævnes kovariansen σ_{12}.

Vi ønsker at finde et tilnærmet udtryk for variansen af funktionen

$$Y = \frac{X_1}{X_2}.$$

Vi starter med at finde de to afledede:

$$\frac{\partial Y}{\partial X_1} = \frac{1}{X_2} \quad \text{og} \quad \frac{\partial Y}{\partial X_2} = -\frac{X_1}{X_2^2}.$$

Værdien af de afledede skal beregnes for μ_1 og μ_2. Jf. (5.9) er et tilnærmet udtryk for variansen givet ved

$$\sigma_Y^2 \approx \left(\frac{1}{\mu_2}\right)^2 \sigma_1^2 + \left(\frac{\mu_1}{\mu_2^2}\right)^2 \sigma_2^2 - 2\frac{\mu_1}{\mu_2^3}\sigma_{12}.$$

Vi kan betragte Y som et estimat af $\mu = \frac{\mu_1}{\mu_2}$ med estimationsvarians σ_Y^2 som angivet ovenfor. Den relative estimationsvarians er forholdet mellem σ_Y^2 og μ^2:

$$\frac{\sigma_Y^2}{\mu^2} = \left(\frac{\sigma_1}{\mu_1}\right)^2 + \left(\frac{\sigma_2}{\mu_2}\right)^2 \frac{2\sigma_{12}}{\mu_1\mu_2}.$$

Øvelse 12
Antag, at X_1 og X_2 er to stokastiske variable som defineret i eksempel 29. Find varians og relativ varians for

$$Y = X_1X_2.$$

I forbindelse med bearbejdningen af landmålingsobservationer vil vi undertiden (bl.a. i forbindelse med udjævninger) skulle bestemme varianser og kovarianser for en række funktioner af de samme stokastiske variable. Vi vil derfor sluttelig udlede en generel fejlforplantningslov med gyldighed for en række differentiable funktioner af n vilkårlige stokastiske variable. Vi starter med at gå videre med eksempel 28 i en mere generel form.

Eksempel 30
Lad X_1 og X_2 være indbyrdes afhængige (korrelerede) med varianserne σ_1^2 og σ_2^2, samt kovariansen σ_{12}. Lad endvidere Y_1 og Y_2 være linearkombinationer af X_1 og X_2:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + b_1 \quad \text{og} \tag{5.10}$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + b_2. \tag{5.11}$$

Vi ønsker at bestemme $\mathbb{V}\text{ar}(Y_1)$, $\mathbb{V}\text{ar}(Y_2)$ samt $\text{Cov}(Y_1, Y_2)$.

Ifølge sætning 12 (side 26) er varianserne for Y_1 og Y_2 givet ved

$$\mathbb{V}\text{ar}(Y_1) = a_{11}^2\sigma_1^2 + a_{12}^2\sigma_2^2 + 2a_{11}a_{12}\sigma_{12} \quad \text{og} \tag{5.12}$$
$$\mathbb{V}\text{ar}(Y_2) = a_{21}^2\sigma_1^2 + a_{22}^2\sigma_2^2 + 2a_{21}a_{22}\sigma_{12}. \tag{5.13}$$

Summen af Y_1 og Y_2 er

$$Y_1 + Y_2 = (a_{11} + a_{21})X_1 + (a_{12} + a_{22})X_2 + b_1 + b_2. \tag{5.14}$$

Variansen for $Y_1 + Y_2$ findes ved at anvende sætning 12 på (5.14)

$$\mathbb{V}\text{ar}(Y_1 + Y_2) = (a_{11} + a_{21})^2\sigma_1^2 + (a_{12} + a_{22})^2\sigma_2^2 + 2(a_{11} + a_{21})(a_{12} + a_{22})\sigma_{12}. \tag{5.15}$$

For at finde kovariansen mellem Y_1 og Y_2 bemærker vi, at sætning 11 siger, at $\mathbb{V}\text{ar}(Y_1 + Y_2) = \mathbb{V}\text{ar}(Y_1) + \mathbb{V}\text{ar}(Y_2) + 2\text{Cov}(Y_1, Y_2)$. Isolerer vi $\text{Cov}(Y_1, Y_2)$, får vi

$$\text{Cov}(Y_1, Y_2) = \frac{1}{2}[\mathbb{V}\text{ar}(Y_1 + Y_2) - \mathbb{V}\text{ar}(Y_1) - \mathbb{V}\text{ar}(Y_2)]$$

Indsættes (5.12), (5.13) og (5.15) får vi

$$\mathrm{Cov}(Y_1, Y_2) = a_{11}a_{21}\sigma_1^2 + a_{12}a_{22}\sigma_2^2 + (a_{11}a_{22} + a_{12}a_{21})\sigma_{12}. \tag{5.16}$$

Vi har således ved hjælp af (5.12), (5.13) og (5.16) fundet generelle udtryk for varianser og kovarianser af to lineære funktioner af to variable.

En alternativ tilgang er at anvende matrix-formuleringen i afsnit 2.9:

$$\mathbf{Y} = \mathbf{AX} + \mathbf{b},$$

hvor

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ og } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Varianser og kovarianser for X_1 og X_2 kan opsummeres i kovariansmatricen $\mathbf{K_X}$ for $\mathbf{X}$:

$$\mathbf{K_X} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix},$$

hvor $\sigma_{12} = \sigma_{21} = \mathrm{Cov}(X_1, X_2)$. Variansen for $\mathbf{Y}$ er nu ifølge sætning 14

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\mathbf{Y}] &= \mathbf{AXA}^T \\ &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11}^2\sigma_1^2 + a_{12}^2\sigma_2^2 + 2a_{11}a_{12}\sigma_{12} & \cdots \\ \cdots & \cdots \end{bmatrix}. \end{aligned}$$

Pga. pladsmangel er kun øverste venstre element i kovarians matricen udregnet. Det ses, at dette element stemmer overens med variansen for Y_1 i (5.12). Det overlades til læseren at verificere, at de tre resterende elementer stemmer overens med (5.13) og (5.16).

Øvelse 13
Udregn varianser og kovarians for Y_1 og Y_2 fra eksempel 28, men forudsæt nu, at X_1 og X_2 er korrelerede, dvs. $\sigma_{12} \neq 0$.

Flere ikke-lineære funktioner

I eksempel 30 var Y_1 og Y_2 lineære funktioner af X_1 og X_2. Vi betragter nu den mere generelle situation, hvor Y_1 og Y_2 er ikke-lineære funktioner af X_1 og X_2:

$$\begin{aligned} Y_1 &= g_1(X_1, X_2) \\ Y_2 &= g_2(X_1, X_2). \end{aligned}$$

Vi antager desuden, at g_1 og g_2 er differentiable funktioner. For at kunne finde varianser og kovarianser for Y_1 og Y_2 lineariserer vi både g_1 og g_2 som i (5.2):

$$Y_1 \approx g_1(\mu_1, \mu_2) + \frac{\partial g_1}{\partial x_1}(X_1 - \mu_1) + \frac{\partial g_1}{\partial x_2}(X_2 - \mu_2)$$
$$Y_2 \approx g_2(\mu_1, \mu_2) + \frac{\partial g_2}{\partial x_1}(X_1 - \mu_1) + \frac{\partial g_2}{\partial x_2}(X_2 - \mu_2).$$

Før vi fortsætter er det bekvemt at omskrive ovenstående lineariseringer af Y_1 og Y_2, så konstante led og variable led (led der involverer X_1 og X_2) er adskilt:

$$Y_1 \approx g_1(\mu_1, \mu_2) - \frac{\partial g_1}{\partial x_1}\mu_1 - \frac{\partial g_1}{\partial x_2}\mu_2 + \frac{\partial g_1}{\partial x_1}X_1 + \frac{\partial g_1}{\partial x_2}X_2 \tag{5.17}$$
$$Y_2 \approx g_2(\mu_1, \mu_2) - \frac{\partial g_2}{\partial x_1}\mu_1 - \frac{\partial g_2}{\partial x_2}\mu_2 + \frac{\partial g_2}{\partial x_1}X_1 + \frac{\partial g_2}{\partial x_2}X_2. \tag{5.18}$$

Bemærk, at (5.17) og (5.18) er på samme form som (5.10) og (5.11), hvor b'erne er givet ved

$$b_1 = g_1(\mu_1, \mu_2) - \frac{\partial g_1}{\partial x_1}\mu_1 - \frac{\partial g_1}{\partial x_2}\mu_2 \quad \text{og} \quad b_2 = g_2(\mu_1, \mu_2) - \frac{\partial g_2}{\partial x_1}\mu_1 - \frac{\partial g_2}{\partial x_2}\mu_2$$

og a'erne er givet ved

$$a_{11} = \frac{\partial g_1}{\partial X_1} \qquad a_{12} = \frac{\partial g_1}{\partial X_2}$$
$$a_{21} = \frac{\partial g_2}{\partial X_1} \qquad a_{22} = \frac{\partial g_2}{\partial X_2}. \tag{5.19}$$

Bemærk, at

$$a_{ij} = \frac{\partial g_i}{\partial x_j}.$$

Ved indsætning i (5.12), (5.13) og (5.16) opnår vi tilnærmede udtryk for varianser og kovarianser.

Bruger vi matrix-formuleringen fra eksempel 30 kan lineariseringerne (5.17) og (5.18) skrives på matrix form som

$$\begin{aligned}\mathbf{Y} &\approx \mathbf{G}(\mathbf{X} - \boldsymbol{\mu}) + \mathbf{g} \\ &= \mathbf{G}\mathbf{X} - \mathbf{G}\boldsymbol{\mu} + \mathbf{g},\end{aligned} \tag{5.20}$$

hvor

$$\mathbf{G} = \begin{bmatrix} \frac{\partial g_1}{\partial X_1} & \frac{\partial g_1}{\partial X_2} \\ \frac{\partial g_2}{\partial X_1} & \frac{\partial g_2}{\partial X_2} \end{bmatrix} \quad , \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{og} \quad \mathbf{g} = \begin{bmatrix} g_1(\mu_1, \mu_2) \\ g_2(\mu_1, \mu_2) \end{bmatrix}.$$

Matricen $\mathbf{G}$ er et eksempel på en såkaldt Jacobi-matrix.

Anvender vi sætning 14 på (5.20), opnår vi, at et tilnærmet udtryk på matrix-form for variansen af $\mathbf{Y}$ er

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\mathbf{Y}] &\approx \mathbb{V}\mathrm{ar}[\mathbf{G}\mathbf{X} - \mathbf{G}\boldsymbol{\mu} + \mathbf{g}] \\ &= \mathbf{G}\mathbb{V}\mathrm{ar}[\mathbf{X}]\mathbf{G}^T \\ &= \mathbf{G}\mathbf{K}_\mathbf{X}\mathbf{G}^T.\end{aligned} \tag{5.21}$$

5.4 Matrix formulering

Det kan vises, at (5.12), (5.13) og (5.16) kan udvides til at omfatte m reelle differentiable funktioner $Y_1, \ldots, Y_m$ af n stokastiske variable

$$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_n) \\ &\vdots \\ Y_m &= g_m(X_1, \ldots, X_n). \end{aligned}$$

Kovariansmatricen for X_i $(i = 1, \ldots, n)$ defineres ved

$$\mathbf{K_X} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}$$

Matricen er kvadratisk og symmetrisk, idet

$$\sigma_{ij} = \sigma_{ji}$$

Vi generaliserer nu matricen af partielle afledede. De partielle afledede (jf. (5.19)) opstilles ligeledes i en matrix, således at første række svarer til de afledede Y_1 med hensyn til $X_1, \ldots, X_n$. Næste række er de afledede af Y_2 osv. indtil sidste række, som er de afledede af Y_m. Værdierne af de afledede skal beregnes i punktet $(\mu_1, \mu_2, \ldots, \mu_n)$ eller for værdier i nærheden af middelværdierne.

Vi får altså en matrix med m rækker og n søjler.

$$\mathbf{G} = \begin{bmatrix} \frac{\partial g_1}{\partial X_1} & \frac{\partial g_1}{\partial X_2} & \cdots & \frac{\partial g_1}{\partial X_n} \\ \frac{\partial g_2}{\partial X_1} & \frac{\partial g_2}{\partial X_2} & \cdots & \frac{\partial g_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial X_1} & \frac{\partial g_m}{\partial X_2} & \cdots & \frac{\partial g_m}{\partial X_n} \end{bmatrix}$$

Sidste skrivemåde kan repræsentere matricen efter udregning af de partielle afledede i eller i nærheden af $(\mu_1, \mu_2, \ldots, \mu_n)$, eller repræsentere koefficientmatricen, hvis $Y_1, \ldots, Y_m$ er lineære funktioner.

Formlerne 5.12, 5.13 og 5.16 og matrixligningen (5.21) kan nu generaliseres:

Definition 14 (Den generelle fejlforplantningslov)
Det tilnærmede udtryk for kovariansmatricen for Y

$$\mathbf{K_Y} \approx \mathbf{G}\mathbf{K_X}\mathbf{G}^T, \tag{5.22}$$

benævnes den generelle fejlforplantningslov. △

Kender vi kovariansmatricen for X'erne, kan vi ved hjælp af (5.22) finde en tilnærmet kovariansmatrix for vilkårlige funktioner af X'erne.

Eksempel 31
Vha. den generelle fejlforplantningslov (5.22) vil vi finde kovariansmatricen for funktionerne

$$Y_1 = 2X_1 - 3X_2$$
$$Y_2 = 3X_1 + 2X_2$$

Da både Y_1 og Y_2 er lineære funktioner af X_1 og X_2 følger det let, at

$$\mathbf{G} = \begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix}.$$

Da vi har to variable X_1 og X_2 i sving har vi

$$\mathbf{K_X} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

Kovariansmatricen for $\mathbf{Y} = (Y_1, Y_2)$ kan nu udregnes:

$$\begin{aligned} \mathbf{K_Y} &= \mathbf{G}\mathbf{K_X}\mathbf{G}^T \\ &= \begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ -3 & 2 \end{bmatrix} \\ &= \begin{bmatrix} (2\sigma_1^2 - 3\sigma_{21}) & (2\sigma_{12} - 3\sigma_2^2) \\ (3\sigma_1^2 + \sigma 12) & (3\sigma_{12} + 2\sigma_2^2) \end{bmatrix} \begin{bmatrix} 2 & 3 \\ -3 & 2 \end{bmatrix} \\ &= \begin{bmatrix} (4\sigma_1^2 + 9\sigma_2^2 - 12\sigma_{12}) & (6\sigma_1^2 - 6\sigma_2^2 - 5\sigma_{12}) \\ (6\sigma_1^2 - 6\sigma_2^2 - 5\sigma_{12}) & (9\sigma_1^2 + 4\sigma_2^2 + 12\sigma_{12}) \end{bmatrix} \end{aligned}$$

Dvs. K_Y er som før.

Af resultatet ses, at hvis $\sigma_1^2 = \sigma_2^2 = \sigma^2$ og $\sigma_{12} = 0$ (X_1 og X_2 er fx. lige gode og uafhængige observationer) fås

$$\mathbf{K_Y} = \begin{bmatrix} 13\sigma^2 & 0 \\ 0 & 13\sigma^2 \end{bmatrix}$$

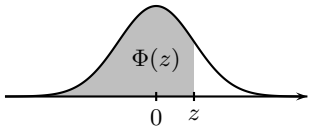
Resultatet indebærer altså, at $\mathrm{Cov}(Y_1, Y_2) = 0$, dvs. Y_1 og Y_2 ikke er korrelerede. Husk, at dette *ikke* betyder, at vi kan konkludere, at Y_1 og Y_2 er uafhængige.$\Diamond$

Øvelse 14
Model og observationsværdier/spredninger fra øvelse 11. Beregn ved hjælp af (5.22) *varianserne for* Y *og* X *samt kovariansmatricen* $\mathrm{Cov}(Y, X)$.

Kapitel 6

Tabeller

Tabel 6.1: Tabel over fordelingsfunktionen $\Phi(z)$ for en standard normalfordelt stokastisk variabel.



z	**,0**	**,1**	**,2**	**,3**	**,4**	**,5**	**,6**	**,7**	**,8**	**,9**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Værdierne i tabellen er fundet vha. software-pakken `R`, `www.r-project.org`