

## KOMPENDIUM OM USIKKERHEDER OG STATISTIK

*Athene Demuth, Børge Svane Nielsen, Kim Lefmann og MekRel Labteamet 2021*

*Version 4.2, Februar 2022*

---

## Introduktion

Fysik beskæftiger sig med den målbare virkelighed, og i det her kompendium kan du læse om, hvordan vi ved hjælp af statistik kan forholde os til den data, som vi kan trække ud af vores verden. Kompendiet introducerer de grundlæggende værktøjer til at bestemme usikkerheder, behandle data og vurdere modeller. Formålet med kompendiet er, at det kan bruges som et opslagsværk og give et indblik i de statistiske koncepter, som skal bruges til behandling af data fra egne eksperimentelle opstillinger i løbet af de første to år af studiet. Der er referencer til andet materiale de steder hvor beviser eller videre læsning kan være relevant. På side 17 er der en opsummerende formelsamling. Hvis du finder fejl eller mangler, så kontakt Athene på mail: glf136@ku.dk. God læselyst :)

## Indhold

<b>1</b>	<b>Intro til usikkerheder og statistik</b>	<b>2</b>
1.1	Betydende cifre . . . . .	2
1.2	Korrekt notation af usikkerheder . . . . .	2
1.3	Statistiske og systematiske usikkerheder . . . . .	2
1.4	Middelværdi fra gentagende målinger . . . . .	3
1.5	Standardafvigelse . . . . .	3
1.6	Usikkerhed på middelværdi . . . . .	4
1.7	Minimering af usikkerheder og $\sqrt{N}$ . . . . .	4
1.8	Normalfordeling . . . . .	4
<b>2</b>	<b>Værktøjer til behandling af data</b>	<b>6</b>
2.1	Enighed blandt eksperimenter . . . . .	6
2.2	Måleusikkerhed og standardafvigelse på samme parameter . . . . .	7
2.3	Usikkerheder med fits . . . . .	7
2.4	Fra usikkerhed i x-aksen til usikkerhed i y-aksen . . . . .	9
2.5	Sammenligning mellem data og modeller - $\chi^2$ . . . . .	10
<b>3</b>	<b>Videregående behandling af data</b>	<b>11</b>
3.1	Ophobningsloven - propagering af usikkerheder . . . . .	11
3.2	Det reducerede chi kvadrat, $\chi^2_{red}$ . . . . .	14
3.3	Fitning med $\chi^2_{red}$ og residualplots . . . . .	15
<b>4</b>	<b>Supplerende materiale</b>	<b>16</b>
<b>5</b>	<b>Formelsamling</b>	<b>17</b>

# 1 Intro til usikkerheder og statistik

Usikkerheder er meget vigtige når vi skal forholde os til en måling. For uden en usikkerhed har et resultat ingen betydning, i og med at vi ikke ved, hvor sikre vi er på resultatet. Hvis vi fx. har målt vægten af en kuffert er det ret relevant om vi bedømte den til at være  $(8.05 \pm 0.02)$  kg eller  $(8 \pm 6)$  kg.

## 1.1 Betydende cifre

Når man noterer en værdi skal man være opmærksom på, hvor mange cifre man sætter på tallet. Medtag aldrig flere cifre, end du er sikker på. De cifre, som man er sikre på, kaldes *betydende cifre* (BC).

### **Eksempel 1.1**

Tallet 34567 har 5 betydende cifre. Hvis vi vil skrive tallet med to betydende cifre, skal det angives  $35 \cdot 10^3$ , og ikke 35000. De efterstillede nuller er ikke betydende cifre, når vi har afrundet. For at gøre det eksplicit hvad der er betydende og ikke-betydende nuller, bruger vi 'scientific notation', hvor nuller angives som potenser.

## 1.2 Korrekt notation af usikkerheder

Ud fra et eksperiment får vi altid en usikkerhed, og det er den usikkerhed, der afgør, hvor mange betydende cifre vi skal afrunde vores resultat til. Den korrekte måde at angive et resultat på er med ét betydende ciffer på usikkerheden. Værdien af resultatet skal derefter afrundes til samme størrelsesorden.

### **Eksempel 1.2**

Ved hjælp af et målebånd er en længde bestemt til 0.12345m, og øjemålet brugt til aflæsningen af målebåndet er vurderet til at være sikkert inde for  $\pm 0.006m$ . Resultatet kan derfor angives som fx.  $(123 \pm 6) \cdot 10^{-3}m$  eller  $(12.3 \pm 0.6)cm$ .

At der kun bør være ét betydende ciffer på usikkerheder kan forklares ved, at man ikke kan kende det efterfølgende betydende ciffer på sin usikkerhed til større præcision end det første ciffer. Det giver derfor ikke mening at inkludere.<sup>1</sup>

## 1.3 Statistiske og systematiske usikkerheder

Vi skelner mellem to forskellige slags usikkerheder, statistiske og systematiske. Begge dele vil være til stede, når vi arbejder i laboratoriet.

*Statistisk usikkerhed* er det, der får vores resultater til at fluktuere tilfældigt omkring den sande værdi. Der vil stort set altid være en statistisk usikkerhed på vores data. Den statistiske usikkerhed på vores resultater kan fx. komme fra vores måleinstrumenter (en køkkenvægt kunne have en statistisk usikkerhed på 0.05g). Når vi tager den samme måling flere gange, vil resultaterne altså ikke være enslydende. Hvis man har en lav statistisk usikkerhed, kalder man resultatet *præcist*.

<sup>1</sup>En undtagelse kan fremkomme, hvis det første ciffer i usikkerheden er 1. I så fald kan det næste betydende ciffer tages med. Nogle tekster og kurser vælger konsekvent at bruge to betydende cifre.

*Systematisk usikkerhed* vil få vores målinger til konsekvent at være forskellig fra den sande værdi. Ofte kaldes de systematiske usikkerheder for *fejlkilder*. Et eksempel på en systematisk usikkerhed kunne være, hvis vores vægt ikke var nulstillet, og derfor systematisk viste en vægt, der var 10g højere end den sande værdi. Eller vi kunne have glemt at tage højde for friktion. Hvis man har lav systematisk usikkerhed, kalder man resultatet for *akkurat*.

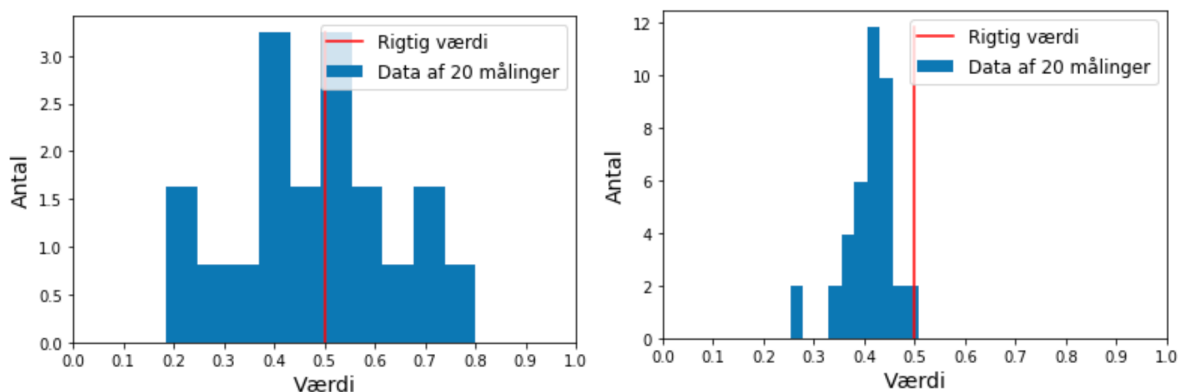


Figure 1: (Venstre) Her ses et eksempel på nogle målinger af den samme opstilling, der har en høj statistisk usikkerhed. (Højre) Her har vi en lavere statistisk usikkerhed, men der er en fejlkilde, som gør at vores data ligger systematisk forskudt fra den sande værdi.

## 1.4 Middelværdi fra gentagende målinger

Flere målinger af den samme størrelse, taget under de samme betingelser, vil give lidt forskellige resultater grundet statistisk usikkerhed. Når vi har sådan et sæt af  $N$  målinger  $[x_1, x_2, x_3 \dots x_N]$ , så vil det bedste bud på den rigtige måleværdi være middelværdien  $\bar{x}$  for datasættet.<sup>2</sup> Estimatet på middelværdien  $\bar{x}$  er defineret

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (1)$$

Usikkerheden på middelværdien,  $\Delta$ , kan udregnes som en statistisk usikkerhed ved at kigge på alle de målinger, der tilsammen gav et estimat af middelværdien.<sup>3</sup>

## 1.5 Standardafvigelse

Når man tager flere målinger af en værdi under de samme betingelser, så vil der naturligt fremkomme en vis spredning af datapunkterne. Til at kvantificere dette definerer vi standardafvigelsen,  $\sigma$ , også kaldet std (standard deviation). Standardafvigelsen afhænger af, hvor langt de individuelle datapunkter gennemsnitligt ligger fra middelværdien  $\bar{x}$ . Hvis alle datapunkter ligger tæt på hinanden, vil  $\sigma$  være lille.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (2)$$

<sup>2</sup>Alle målinger skal have samme usikkerhed, som gerne må være 0, for at man kan benytte udtrykket. Ellers skal man tage et vægtet gennemsnit i forhold til usikkerhederne.

<sup>3</sup>Læs om hvordan du udregner middelværdi og usikkerhed i sektion 4.1.2 af Pythonnoten.

Faktoren  $1/\sqrt{N}$  i udtrykket tager højde for, at jo flere målinger vi har, jo større værdi får vores sum. På grund af den faktor, så vil standardafvigelsen tendere mod en bestemt værdi, uafhængigt af antal målinger. Så det, der står under kvadratroden, er en form for gennemsnitlig afvigelse fra hvert punkt til middelværdien. Størstedelen af datapunkterne vil ligge i intervallet  $\bar{x} \pm \sigma$ .

## 1.6 Usikkerhed på middelværdi

Usikkerheden på selve middelværdien er givet  $\Delta = \sigma / \sqrt{N}$ . Det er  $\Delta$  vi noterer, når vi angiver en usikkerhed på den beregnede middelværdi af et datasæt.<sup>4</sup> Når vi angiver usikkerheder i plots, hvor hvert punkt består af et gennemsnit af flere målinger, er det usikkerheden  $\Delta$  vi angiver til errorbaren.<sup>5</sup>

### Eksempel 1.3

Vægten af en kugle bliver målt til  $[16, 17, 25, 21]g$  ved fire forsøg. Gennemsnitsværdien for datasættet er  $79/4 = 19.75$  og standardafvigelsen udregnes til at være  $\sqrt{50.75/4} = 3.562$ . Så usikkerheden bliver  $3.562/\sqrt{4} = 1.781$  og resultatet angives  $(20 \pm 2)g$ .

## 1.7 Minimering af usikkerheder og $\sqrt{N}$

Fra udtrykket for  $\Delta$  kan vi se at hvis vi tager flere målinger, altså gør  $N$  større, så vil vores usikkerhed på middelværdien falde. Helt specifikt vil usikkerheden falde proportionalt  $1/\sqrt{N}$ . Laver vi vores måling en faktor 100 gange mere (fx fra 3 gange til 300 gange), så ville usikkerheden på middelværdien reduceres med en faktor 10. Vi kan således opnå mindre usikkerhed ved at gentage målingerne. En anden metode til at minimere usikkerhed er at bruge mere præcist udstyr, for hvis usikkerheden på den enkelte måling bliver mindre, så reduceres fluktuationerne omkring middelværdien, hvilket mindsker standardafvigelsen og herved usikkerheden på middelværdien.

## 1.8 Normalfordeling

Hvis vi tager  $N$  målinger af en værdi med spredningen  $\sigma$ , og plotter data i et histogram, så vil histogrammet ofte tage form som en normalfordeling, når  $N$  øges. Se eksempelvis figur 2. Funktionen for en normalfordeling (en Gauss) er givet

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (3)$$

Hvor  $\sigma$  er standardafvigelsen og  $\mu$  er et udtryk for middelværdien,  $\bar{x}$ .<sup>6</sup> Det, at vores histogram vil begynde at ligne en normalfordeling, når vi øger antallet af eksperimenter er et udtryk for *the Central Limit Theorem*, CLT: Når vi har et datasæt, hvor hvert datapunkt er summen af  $N$  uafhængige variable (uafhængige betyder ingen systematisk usikkerhed - alle variable tages tilfældigt fra den samme fordeling), og lader  $N \rightarrow \infty$ , så vil datasættet følge en normalfordeling.<sup>7</sup>

<sup>4</sup>Se Barlow kapitel 4.2.1 for mere om usikkerhed på middelværdi (standard error on the mean). Bemærk, at andre tekstbøger vælger at bruge  $\delta$  eller  $\sigma$  om usikkerheden på en middelværdi.

<sup>5</sup>Læs mere om hvordan du laver errorbars i sektion 4.2.1 af Pythonnoten.

<sup>6</sup>I det her kompendium har vi approksimeret  $\bar{x}$  til at være den sande værdi,  $\mu$ . I virkeligheden er  $\bar{x}$  blot et estimat ud fra data, hvilket man bruger ét sæt information på, og derfor burde  $N - 1$  bruges i stedet for  $N$  til at udregne  $\sigma$ . Læs Barlow kapitel 2.4.2 for mere om Bessel korrektionen og std.

<sup>7</sup>Se Barlow kapitel 3.4 for mere om normalfordelingen.

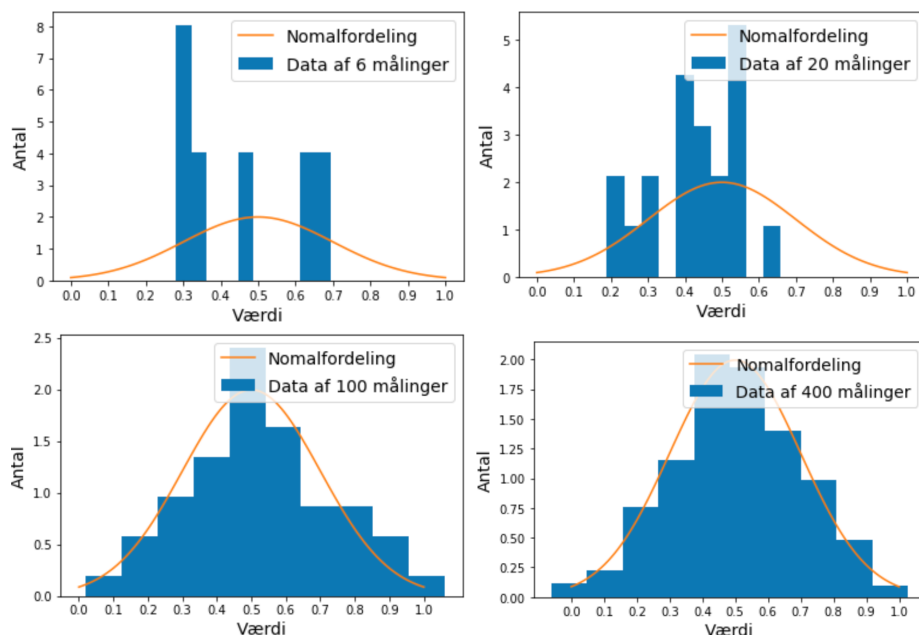


Figure 2: Histogrammer for henholdsvis 6, 20, 100 og 400 målinger af samme værdi. Funktionerne viser den sande fordeling af data, som her er en normalfordelingen på grund af CLT, da usikkerheden på data stammer fra mange forskellige bidrag. Det ses, at allerede ved 100 målinger ligner data normalfordelingen ret godt. Som tommelfingerregel bør antallet af bins i et histogram være lig  $\sqrt{N}$ .

Navnet på sætningen skyldes, at vi skal lade  $N$  blive meget stor (fluktuationerne i data skal stamme fra mange forskellige bidrag), før data følger normalfordelingen.<sup>8</sup> I figur 3 ses, at der er samme andel datapunkter inden for intervallet  $\bar{x} \pm \sigma$ , selvom  $\sigma$  er forskellige. Dette er fundamentalt for alle normalfordelinger, og helt specifikt ligger 68.27% af data i det interval. I intervallet  $\bar{x} \pm 2\sigma$  ligger 95.45% af data, og inde for  $\bar{x} \pm 3\sigma$  vil der være 99.73%.

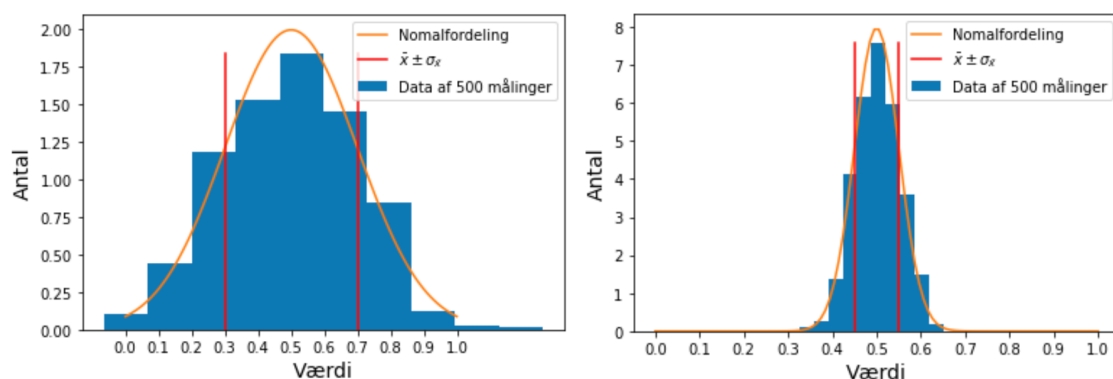


Figure 3: Histogrammer for to datasæt der begge har middelværdi 0.5. Til venstre er  $\sigma = 0.2$ , og til højre er  $\sigma = 0.05$ . De to normalfordelinger er normaliserede, så arealet under kurven er 1. Der er ca. samme andel datapunkter inden for de røde linjer!

<sup>8</sup>Dette er forklaringen på, hvorfor usikkerheder fra eksperimenter er gaussiske - fluktuationen i data fra  $f_x$  en vægt kommer fra mange forskellige faktorer, såsom vind, placering luftfugtighed og spænding fra strømgiver. Tilsammen vil de gøre fluktuationen gaussisk og data bliver normalfordelt.

## 2 Værktøjer til behandling af data

På baggrund af de grundlæggende statistiske begreber kan vi nu begynde at analysere vores data kvantitativt. Blandt andet vil vi se på, hvordan man kan bedømme om eksperimenter er enige, og hvordan forskellige modeller kan sammenlignes med data.

### 2.1 Enighed blandt eksperimenter

Ofte ønsker vi med vores eksperiment at måle en værdi, og sammenligne med værdien fra et andet eksperiment, eller en tabelværdi. Vi har således to værdier  $A \pm \Delta_A$  og  $B \pm \Delta_B$ . Nu kan vi tage differensen mellem værdierne,  $D$  og udregne usikkerheden på værdiernes afvigelse,  $\Delta_D$ , ved

$$D = |A - B| \quad \Delta_D = \sqrt{\Delta_A^2 + \Delta_B^2} \quad (4)$$

Herfra definerer vi afstanden mellem værdierne i enhed af usikkerheden som  $z = D/\Delta_D$ . Når  $z \leq 1$ , altså når afvigelsen er mindre end usikkerheden, siger vi, at eksperimenterne er *enige* inde for et '1 $\Delta$  interval'. I stedet for at sammenligne to eksperimenter kan man vælge at sammenligne et resultat med en tabelværdi ved hjælp af  $z$ -værdien. Tabelværdiens usikkerhed kan være 0.

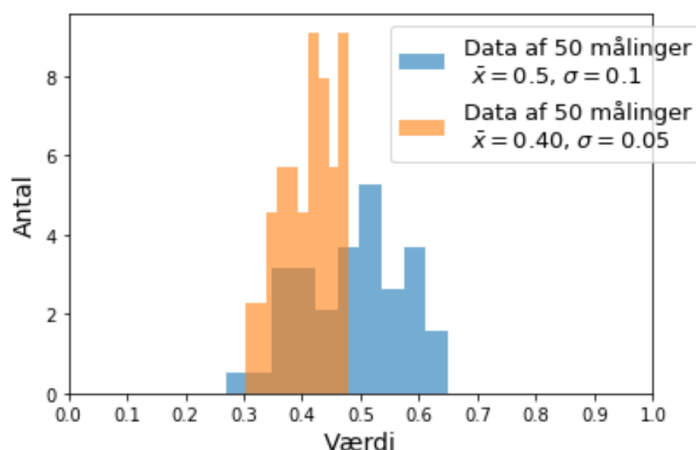


Figure 4: Histogrammer med data fra to forskellige eksperimenter, der måler den samme værdi. Usikkerheden på afstanden mellem de to middelværdier er  $\sqrt{0.1^2 + 0.05^2} / \sqrt{50} \approx 0.02$ , hvilket giver en  $z$ -værdi på 5. Så eksperimenterne er ret uenige om middelværdien.

#### Eksempel 2.1

Vi har målt svingningstiden af et pendul ved to forskellige snorlængder og får henholdsvis  $(1.4 \pm 0.4)$  s og  $(2.00 \pm 0.09)$  s. Vi udregner nu afvigelsen  $|1.4\text{ s} - 2.00\text{ s}| = 0.6\text{ s}$  og usikkerheden på afvigelsen  $\sqrt{0.4^2 + 0.09^2} = 0.41\text{ s} \approx 0.4\text{ s}$ , så  $z = 1.5$ . Altså ser vi, at de to eksperimenter ikke er enige om svingningstiden, inden for et interval på  $\pm 1\Delta$ . Men de er enige inde for intervallet  $\pm 2\Delta$ . Hvis man vil være mere sikker på, om de to eksperimenter er enige, bør man tage flere målinger. Er et sæt målinger enige, så bør de to værdier rykke tættere på hinanden, efterhånden som usikkerhederne vil falder med flere målinger.

Oftentimes kigger vi på et større interval end én gang usikkerheden. Det gælder særligt, hvis vi vil undersøge, om to resultater *ikke* er enige. Higgs partiklen blev fx. fundet ved at data lå mere end  $5\Delta$  væk fra den værdi man forventede at finde uden eksistensen af Higgs'en! Det svarer til en 99.99995% sandsynlighed for, at det ikke blot var en statistisk fluktuation.

## 2.2 Måleusikkerhed og standardafvigelse på samme parameter

I nogle tilfælde har vi både estimeret en måleusikkerhed og har taget målingen flere gange, hvilket giver en standardafvigelse. Det er *altid* den udregnede statistiske usikkerhed, der bør angives. Men det er også relevant at sammenligne måleusikkerheden med den statistiske usikkerhed. Hvis det kun var måleusikkerheden, der gav anledning til afvigelserne, så burde de to usikkerheder være ens. Ved at sammenligne måleusikkerhed og den udregnede statistiske usikkerhed kan vi vurdere, hvor usikkerhederne kommer fra.

### Eksempel 2.2

En måleusikkerhed kommer meget ofte fra en standardafvigelse. Producenten af fx. en vægt eller en laser udregner deres usikkerhed ud fra måling af standardafvigelse. Og hvis vi aflæser en digital værdi på et voltmeter til at være 1.5 V, så har vi en usikkerhed på  $\pm 0.05$  V, fordi resultatet er afrundet. Afrundingen kommer fra antallet af betydende cifre, der er fundet ud fra usikkerheden, der er fundet ved en standardafvigelse. Og når vi giver en aflæsnings usikkerhed, kommer det netop fra en vurdering af et interval, hvori vi er nogenlunde sikre på, at vi har ramt rigtigt - altså et kvalitativt bud på en standardafvigelse. Så på den måde kan man sige at alle usikkerheder kommer fra standardafvigelser! Stort set alle de usikkerheder vi beskæftiger os med er gaussiske og vores data vil følge normalfordelingen.

## 2.3 Usikkerheder med fits

I laboratoriet undersøger vi ofte, hvordan én parameter afhænger af en anden parameter, og vi kender den teoretiske funktionsforskrift, som forventes at beskrive data. Et godt værktøj er *curve\_fit*, som er en funktion i Python under *scipy.optimize*. Ved hjælp af den kan vi fitte data efter modellen og inkludere usikkerhed i y-aksens retning for hvert datapunkt.

Når vi bruger værktøjet fitter vi efter en funktion, der vil have et antal frie parametre<sup>9</sup>, og vi vil få outputtet de værdier for de  $n$  frie parametre, der passer bedst til data. Desuden får vi en covariansmatrice, der i diagonalen indeholder kvadratet på usikkerheden for hver af parametrene.<sup>10</sup> Ved at tage kvadratroden af diagonalen af covariansmatricen, får man usikkerheden på hver af modellens frie parametre.

<sup>9</sup>Frie parametre er antallet af ukendte værdier i funktionsforskriften. Et lineært fit  $ax + b$  har to frie parametre, som man kan fitte efter, nemlig  $a$  og  $b$ .

<sup>10</sup>Covariansmatricen er en symmetrisk matrix, der udover at indeholde kvadratet på usikkerhederne i dens diagonal, også indeholder covariansen mellem hver af parametrene indbyrdes. En covarians er et mål for, hvor meget variablene hænger sammen. Hvis den er høj, så vil en lille ændring i den ene parameter føre til en stor variation i den anden, og omvendt.



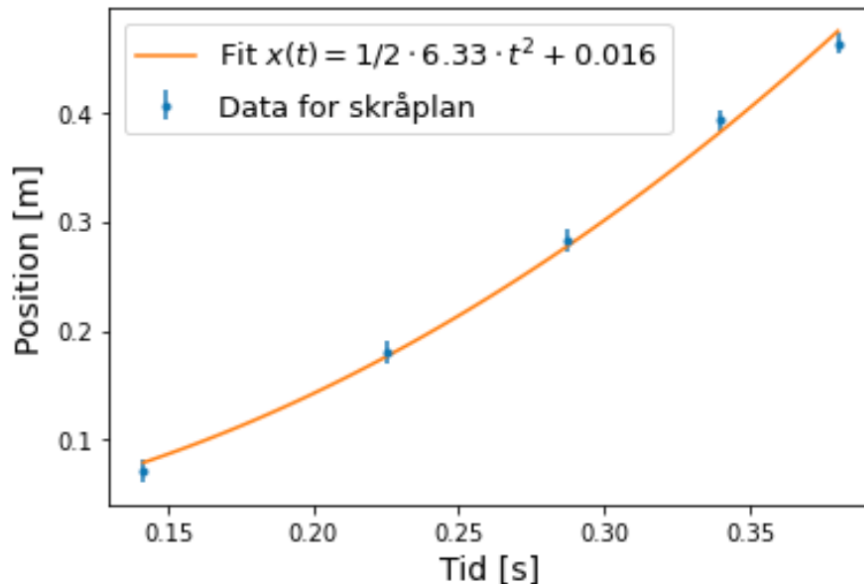


Figure 5: Data for et forsøg med et skråplan. Position og tidspunkt er noteret for 5 punkter. Usikkerheden for positionen på hver af punkterne er  $\pm 1\text{ cm}$ . Ved at fitte til data får vi en acceleration på  $(6.3 \pm 0.2) \text{ m/s}^2$  og en skæringsværdi på  $(0.016 \pm 0.009) \text{ m}$ .

Usikkerheden, som vi kan angive til `curve_fit`, er et array med usikkerheden på hvert datapunkt. Usikkerheden kan sagtens være forskellig for hvert datapunkt. I det tilfælde vil størrelsen af fit parametrene afhænge af forholdet mellem usikkerhederne. Hvis et datapunkt har lav usikkerhed, så vil punktet blive vægtet højere, i forhold til hvor `curve_fit` placerer fittet, end hvis det havde en stor usikkerhed. Usikkerheden på hver af parametrene vil også blive påvirket. Hvis alle datapunkter har samme usikkerhed, så vil tilføjelsen af usikkerheder ikke påvirke størrelsen af værdien for fit parametrene, men vil ændre parametrene usikkerhed.<sup>11</sup>

### Eksempel 2.3

Fittet for data med skråplanet, der er vist i figur 5, giver en skæringsværdi på  $(1.6 \pm 0.9) \text{ cm}$ . Ud fra det kan vi konkludere, at vi måske har placeret kuglen et lille stykke væk fra det sted, som vi definere som  $x = 0 \text{ cm}$ . Afstanden mellem den teoretiske værdi, 0, og den eksperimentelle værdi, i enhed af usikkerheden er  $z = 1.8$ . Fra det kan vi sige, at der kan have været en lille systematisk usikkerhed, der har forskudt alle vores datapunkter.

Når man fitter er det *altid* en god ide at inkludere en skæringsværdi  $y_0$ . Også hvis man fitter efter en teoretiske model, der ikke har en skæringsværdi (fx. kunne det være en proportional model). Når man derefter har fundet fit-parametrene, så kan vi vurdere, hvor forskellig værdien er fra nul i forhold til værdiens usikkerhed. Hvis skæringsværdien er markant forskellig fra nul er det ofte noget, der kan lægge en god grund for diskussion af fejlkilder.

<sup>11</sup>Læs mere om hvordan `curve_fit` bruges i Pythonnoten. HUSK at sætte `absolute_sigma= True`. Ellers bliver usikkerhederne på fit parametrene forkert skaleret.



## 2.4 Fra usikkerhed i x-aksen til usikkerhed i y-aksen

Datapunkter har som regel både en usikkerhed på y-aksen OG på x-aksen. I de tidligere sektioner har vi kun betragtet usikkerheden på y-aksen. Det skyldes bl.a. at *curve\_fit* kun kan inkludere usikkerhed i y-aksens retning. For at kunne fitte korrekt bør usikkerheden i x-aksen inkluderes. Derfor skal vi have et mål for den indvirkning en x-usikkerhed vil have på usikkerheden i y-aksen. Det kan vi gøre via størrelsen af hældningen på den model, vi har fundet ud af beskriver data bedst.<sup>12</sup>

$$\Delta_{y, \text{ fra } x} = \left| \frac{df}{dx} \right| \cdot \Delta_x \quad (5)$$

Når man bruger ligning (5) skal man differentiere sin model, indsætte sine parametre og værdien for det punkt, som man ønsker at finde den nye usikkerhed på, og til sidst gange med usikkerheden i x-aksen.

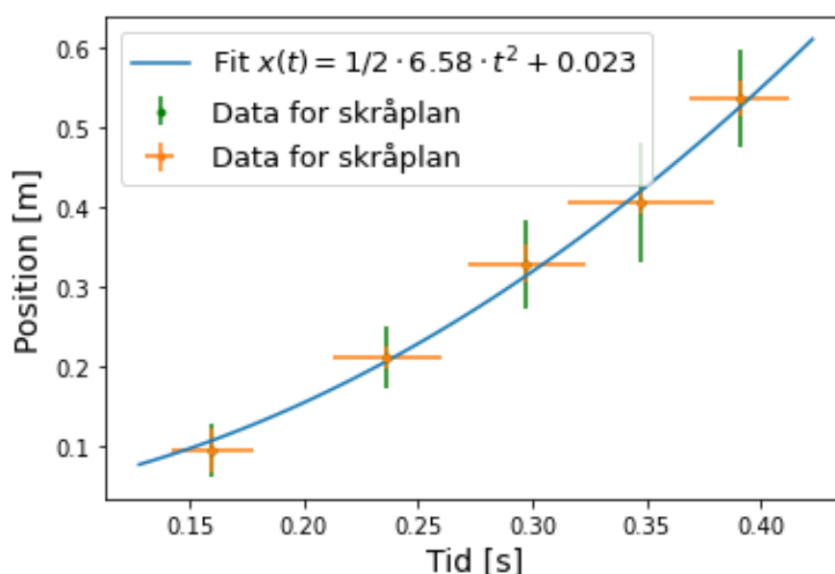


Figure 6: Data for et forsøg med et skråplan. Position og tidspunkt er noteret for 5 punkter ved at måle i håndsbredden og bruge et lomme ur, hvilket giver høje usikkerheder. De grønne errorbars viser den totale usikkerhed, propageret til at være udelukkende i y-aksen, og de orange viser de gamle usikkerheder.

Figur 6 viser et forsøg, hvor data har usikkerhed i både x-aksen og y-aksen, og hvor det derfor er nødvendigt at omregne x-usikkerheden til en y-usikkerhed. For at finde den totale usikkerhed i y-aksens retning lægges de to usikkerheder sammen kvadreret.

$$\Delta_{y, \text{ total}} = \sqrt{\Delta_y^2 + \Delta_{y, \text{ fra } x}^2} \quad (6)$$

Nu har vi reduceret den totale usikkerhed til y-aksens retning. Fra formel (5) ser vi at forholdet mellem den gamle og den nye usikkerhed afhænger af hældningen. Hvis modellen har en høj hældning, så vil en lille variation i x-parameteren give en stor variation i y-parameteren. Derved får x- usikkerheden en større indflydelse.

<sup>12</sup>Se udledningen i Barlow, side 56.

Efter at have fundet den totale usikkerhed, kan vi udføre vores *curve\_fit* igen. På grund af de nye usikkerheder vil punkterne nemlig nu være vægtet lidt forskelligt, hvilket måske vil give anledning til at parametrene ændres lidt. Bemærk, at da vi differentierede modellen og indsatte, så brugte vi de gamle parametre. I princippet kunne vi nu finde et bedre mål for størrelsen af  $\Delta_{y, \text{ fra } x}$ , ved at indsætte de nye parametre, og så finde igen finde  $\Delta_{y, \text{ total}}$  og så til sidst bruge *curve\_fit* for tredje gang. Men det er ofte alt for omstændigt i forhold til hvor lille en korrektion af fit parametrene det vil give.

## 2.5 Sammenligning mellem data og modeller - $\chi^2$

I fysik arbejder vi ofte med kendte modeller, og vil undersøge om de passer til virkeligheden. Derfor er det relevant at have et kvantitativt mål for, hvor godt en model passer til data. Det gør vi ved at kigge på forskellen mellem hvert datapunkt, og den værdi, som modellen forudsiger. Afvigelsen skal ses i forhold til størrelsen af usikkerheden. Præcis som i sektion 2.2 *Enighed blandt eksperimenter*. Nu er usikkerheden på modellen konsekvent 0, men på samme måde kigger vi på om  $|y_{\text{data}} - y_{\text{model}}| / \Delta_{y, \text{ data}}$  er større eller mindre end 1. Da vi kigger på helheden af modellen, summer vi alle de  $N$  datapunkters afvigelse. Desuden tager vi kvadratet på både afvigelsen og usikkerheden. Denne størrelse kalder man et chi-kvadrat,  $\chi^2$ .

$$\chi^2 = \sum_{i=1}^N \frac{(y_{\text{data},i} - y_{\text{model},i})^2}{\Delta y_{\text{data},i}^2} \quad (7)$$

Hvor  $y_{\text{model},i}$  er værdien af modellen taget til punktet  $x_{\text{data},i}$ , altså  $f(x_{\text{data},i})$ . Jo mindre  $\chi^2$ , des bedre overensstemmelse mellem model og data. Hvis en model er helt kendt (altså *ikke* fittet), og  $\chi^2$  for modellen er i størrelsesorden af antallet af datapunkter,  $N$ , så passer modellen godt til data. Får man en værdi der er lavere end  $N$ , så er de usikkerheder man har brugt til data nok overvurderede.

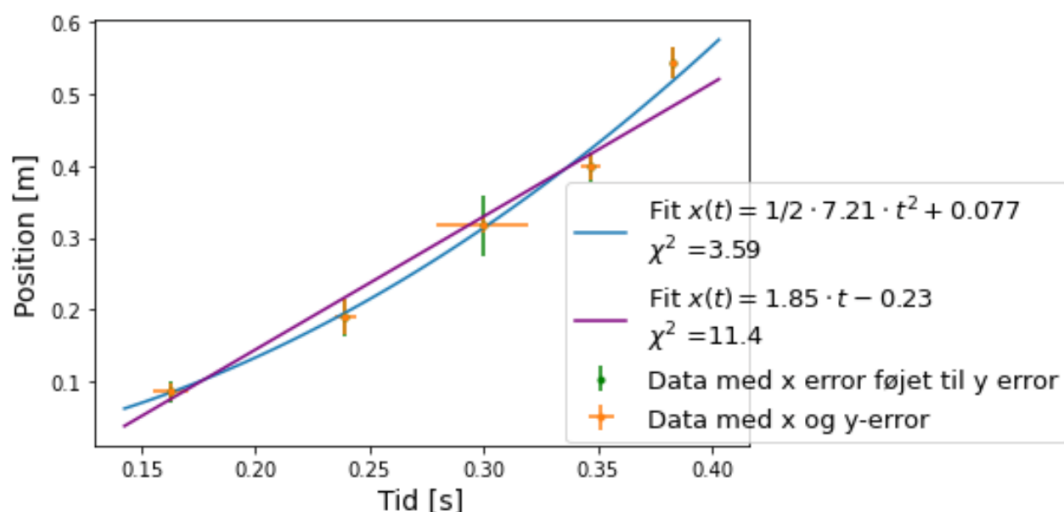


Figure 7: Data for et forsøg med et skråplan. Her har vi sammenlignet data med et andengradspolynomie og en ret linje. Som forventet giver andengradspolynomiet et lavere chi-kvadrat. Ved at omregne fra x- og y-usikkerhed til en samlet y usikkerhed fås et  $\chi^2$  på 2.94 og 8.90. Det skyldes, at usikkerheden, der står i nævneren i formel (7), nu er større.

Bemærk, at der kun indgår usikkerhed i y-retningen i ligning (7). Hvis der også er usikkerheder i x-aksens retning, så skal man omregne fra x-usikkerhed til y-usikkerhed, som beskrevet i sektion 2.4 *Fra usikkerhed i x-aksen til usikkerhed i y-aksen*. Figur 7 viser et datasæt, der er plottet med to forskellige modeller. Ved at udregne  $\chi^2$  for forskellige modeller, kan vi vurdere, hvilken model der passer bedst til data.

#### Eksempel 2.4

*Hvorfor bør vi forvente at et godt chi kvadrat giver omkring N? Hvis vi har den perfekte model til data, så vil den eneste afvigelse mellem modellen og datapunkterne komme fra den statistiske usikkerhed på data. For vores model kan ikke være mere perfekt end vores data. Ud fra det kan vi konkludere, at afvigelsen mellem den perfekte model og datapunkterne vil tilsvare usikkerheden på data.*

$$\sum_{i=1}^N (y_{data,i} - y_{model,i})^2 = \sum_{i=1}^N (\Delta y_{data,i})^2$$

*Lad os antage at usikkerheden er det samme for hvert punkt. Det betyder at højresiden blot giver  $N \cdot \Delta y_{data}^2$ . Ud fra det kan vi se at standardafvigelsen mellem model og datapunkter faktisk er det samme som usikkerheden på selve data! Da vi nu har antaget at usikkerheden er konstant, kan vi også trække den ud fra summen i chi kvadratet. Desuden kan vi indsætte udtrykket for summen af afvigelsen mellem model og data.*

$$\chi^2 = \frac{1}{\Delta y_{data}^2} \sum_{i=1}^N (y_{data,i} - y_{model,i})^2 = \frac{1}{\Delta y_{data}^2} N \cdot \Delta y_{data}^2 = N$$

*Altså ser vi, at hvis vi har den perfekte model til data, så er  $\chi^2 \approx N$ . I tilfældet af at chi kvadratet er mindre end N, så må der være en fejl i vores fortolkning af data. For modellen kan ikke passe bedre til*

## 3 Videregående behandling af data

Her kommer de sidste brikker vi skal bruge, for at kunne lave en gennemgående databehandling. Mange af jer har nok undret jer over, hvad vi skal gøre, hvis vi har usikkerheder på flere parametre, og her kommer en formel, der samler det hele! Desuden kigger vi på, hvad der sker med chi-kvadratet når vi fitter en model.

### 3.1 Ophobningsloven - propagering af usikkerheder

Eksperimenterne vi laver inddrager ofte flere end parameter med usikkerhed. derfor har vi brug for et værktøj, der kan samle alle usikkerhederne.<sup>13</sup> For at kunne samle alle usikkerhederne skal man have en model  $f$ , der beskriver relationen mellem parametrene. Ophobningsloven er det værktøj, som vi kan bruge til at finde den propagerede usikkerhed på  $f$ .

<sup>13</sup>Læs mere om ophobningsloven i sektion 4.3.2 af Barlow.

$$\Delta f = \sqrt{\sum_{i=1}^n \left( \frac{\partial f}{\partial \alpha_i} \cdot \Delta \alpha_i \right)^2} \quad (8)$$

Hvor  $\alpha_i$  er de  $n$  parametre, der er i modellen.<sup>14</sup> Hvis  $f$  fx var bevægelsesligningen for et frit fald, så ville den have tyngdeaccelerationen, starthastighed og startposition som sine tre parametre. Bemærk, at formel (5) fra sektion 2.4 *Fra usikkerhed i x-aksen til usikkerhed i y-aksen* præcis er ophobningsloven med én parameter!<sup>15</sup>

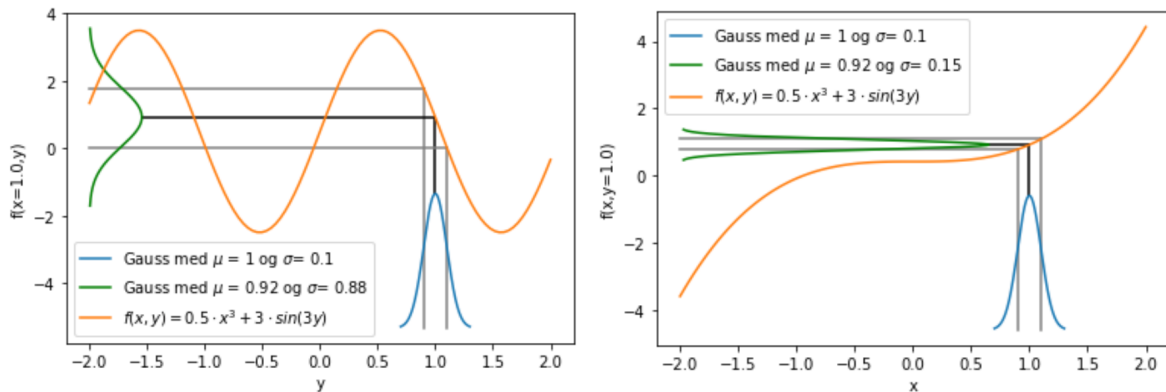


Figure 8: Illustration af ophobningsloven for en funktion  $f(x,y) = 1/2 \cdot x^3 + 3 \sin(3y)$ . Parametrene  $x$  og  $y$  bliver begge målt til at være  $1.0 \pm 0.1$ . Men fordi  $x$  og  $y$  har forskellige relationer til  $f$ , vil deres usikkerheder supplere forskelligt til usikkerheden på  $f$ . I plottet markerer de sorte linjer parameterens værdi, og funktionen  $f$  taget til parameteren, mens at de lodrette grå linjer viser intervallerne  $x \pm \Delta x$  og  $y \pm \Delta y$ . De vandrette grå linjer viser den propagerede usikkerhed på  $f$  fra hver parameter. Denne sikkerhed på  $f$  svarer præcis til den værdi, som man ville have fået ved at udregne  $\partial f / \partial x \cdot \Delta x$  og  $\partial f / \partial y \cdot \Delta y$ . Fra plottet til venstre ses tydeligt, at hvis funktionen  $f(1.0, y)$  har en høj hældning omkring den værdi  $y$ , som man måler sin parameter til at være, så vil  $\Delta y$  have en stor indflydelse på  $\Delta f$ . Hvis  $f(x, 1.0)$  derimod, som i plottet til højre har en lav hældning omkring  $x$ , så vil  $\Delta x$  have en meget lille indflydelse på  $\Delta f$ . Den totale usikkerhed på  $f(1.0, 1.0)$  kan udregnes ved ophobningsloven til at være  $\sqrt{0.88^2 + 0.15^2} = 0.9$ .

I ophobningsloven indgår en sum. Det betyder, at vi skal forholde os til alle de  $n$  parametre, som vores model indeholder. Når vi bruger ophobningsloven, så vil de usikkerheder vi indsætter komme fra en blanding af måleusikkerheder, usikkerheder bestemt ved at bruge  $\sigma / \sqrt{N}$  og usikkerheder fundet fra fit.

Desuden indeholder den de partielt afledede af hver af parametrene. Så hvis modellen afhænger meget af en bestemt parameter - dvs. at den partielt afledede er 'stejl', så vil parameteren betyde meget for den samlede usikkerhed. Det er praktisk at kigge på leddene i ophobningsloven hver for sig, og derfra vurdere, hvilken parameter der har størst indflydelse på usikkerheden. Jo større usikkerhed et parameter led har, des mere vil den supplere til den samlede usikkerhed på  $f$ . Se eksempelvis figur 8.

<sup>14</sup>Man kan tænke på ophobningsloven som en gradient i parameterrummet, der er skaleret i forhold til størrelsen af usikkerheden på hver af parametrene.

<sup>15</sup>Begge disse formler er approksimationer, og gælder kun såfremt at usikkerhederne på parametrene ikke er korrelerede,  $f''(\alpha)$  ikke er for stejl omkring funktionsværdien, relativt til usikkerheden og at usikkerheden på parameteren  $\alpha$  er Gaussisk. Disse undtagelser kommer vi dog sjældent ud for i laboratoriet.

**Eksempel 3.1**

Hvis vi vil lægge to tal sammen, der hver har en usikkerhed bruger vi naturligvis ophobningsloven. Vores udtryk kan vi kalde  $D = A + B$ . Og nu kan vi skrive ophobningsloven op på  $\Delta D$

$$\Delta D = \sqrt{\left(\frac{\partial D}{\partial A}\right)^2 \Delta_A^2 + \left(\frac{\partial D}{\partial B}\right)^2 \Delta_B^2}$$

Hvilket giver os

$$\Delta D = \sqrt{\Delta_A^2 + \Delta_B^2}$$

Som præcis er den formel, som vi brugte tidligere i ligning (4), når vi tog en sum eller differens af to målinger! Ved ophobningsloven kan vi også udlede, at hvis vi har en konstant gange en variabel, så er usikkerheden blot givet

$$\Delta f = k \cdot \Delta x, \quad \text{hvor} \quad f(x) = kx$$

Vi kan altså bruge ophobningsloven til at genfinde de formler, som vi tidligere har brugt til at udregne usikkerheder.

Ophobningsloven givet ved formel (8) antager, at funktionen kan approximeres nogenlunde ved en Taylor ekspansion til første orden omkring vores parameter, i et interval i størrelsesorden af usikkerheden på parameteren. Derved kan vi bruge hældningen til at estimere relationen mellem usikkerheden parameteren og på vores værdi  $f$ . Men hvis hældningen ændrer sig brat og asymmetrisk i intervallet  $x \pm \Delta x$ , så bryder approximationen sammen og ophobningsloven er ikke længere helt gyldig.

**Eksempel 3.2**

Ved måling af hastigheden på en bil tager vi både en længdemåling og en tidsmåling. Det giver resultaterne  $(20.0 \pm 0.5)m$  og  $(1.0 \pm 0.1)s$ . Nu skal vi finde den samlede usikkerhed på hastigheden,  $v = \text{distance} / \text{tid}$ . Først differentierer vi  $v$  partielt i forhold til henholdsvis tiden og distancen:  $\partial v / \partial d = 1/t$  og  $\partial v / \partial t = -d/t^2$ . Det kan vi nu indsætte i ophobningsloven

$$\Delta v = \sqrt{\left(\frac{1}{t}\right)^2 \cdot \Delta_d^2 + \left(\frac{-d}{t^2}\right)^2 \cdot \Delta_t^2}$$

Nu udregnes de to led under kvadratroden hver for sig:

$$\left(\frac{1}{1.0s} \cdot 0.5m\right)^2 = 0.25m^2/s^2$$

$$\left(\frac{20.0m}{(1.0s)^2} \cdot 0.2s\right)^2 = 4m^2/s^2$$

Den totale usikkerhed på hastigheden bliver derved

$$\Delta v = \sqrt{0.25m^2/s^2 + 4m^2/s^2} = 2.06m/s$$

Og hastigheden på bilen er  $20.0\text{m}/1.0\text{s} = 20\text{m/s} = 72\text{km/t}$ . For at få usikkerheden i enheden  $\text{km/t}$  bruger vi blot  $\Delta v_{\text{km/t}} = 3.6 \cdot \Delta v_{\text{m/s}} = 7.4\text{km/t}$ . Så vores resultat på bilens fart bliver  $(72 \pm 7)\text{km/t}$ . Fra udregning af hver af leddene i ophobningsloven kan vi se at det klart er usikkerheden på tiden, der supplerer mest til usikkerheden på hastigheden. Så hvis vi vil forbedre målingen af bilens fart, bør tidsmålingen optimeres.

### 3.2 Det reducerede chi kvadrat, $\chi_{\text{red}}^2$

Når vi udfører et fit af data, enten ved håndkraft eller via *curve\_fit*, så finder man faktisk den model, der minimerer chi kvadratet mest muligt. I det introducerende afsnit til chi kvadrater, sektion 2.5 *Sammenligning mellem modeller og data* kiggede vi hvad et godt chi kvadrat var i tilfældet af en kendt model. Det betød, at så snart vi havde taget vores data, så var  $\chi^2$  fastlagt. Men når vi fitter ved at minimere chi kvadratet, er  $\chi^2$  en funktion af parametrene. Det betyder, at vi mister et antal frihedsgrader,  $n$  - én for hver parameter vi fitter efter. Derfor kan vi ikke sammenligne direkte mellem  $\chi^2$  og antallet af datapunkter,  $N$ . I stedet skal vi sammenligne med antallet af frihedsgrader,  $N_{\text{dof}} = N - n$ . Det er praktisk at definere en ny værdi, det reducerede chi kvadrat  $\chi_{\text{red}}^2$ , der giver et mål for, hvor godt en fittet model passer til data.<sup>16</sup>

$$\chi_{\text{red}}^2 = \frac{\chi^2}{N_{\text{dof}}} = \frac{1}{N_{\text{dof}}} \sum_{i=1}^N \frac{(y_{\text{data},i} - y_{\text{model}})^2}{\Delta y_{\text{data},i}^2} \quad (9)$$

Hvis modellen er god, vil  $\chi_{\text{red}}^2$  være tæt på 1. Og jo større det reducerede chi kvadrat er, des værre er modellen. Er  $\chi_{\text{red}}^2$  mindre end 1, så har man sandsynligvis overvurderet usikkerhederne (se eksempel 2.3). Udtrykket kan også bruges, hvis man ikke har nogle fit parametre. Så har vi samme situation som i sektion 2.5 *Sammenligning mellem modeller og data*. Som regel bruger man slet ikke  $\chi^2$ , fordi  $\chi_{\text{red}}^2$  er nemmere at forholde sig til, og man behøver ikke at have kendskab til antallet af datapunkter eller fit parametre, for at kunne vurdere modellen ud fra  $\chi_{\text{red}}^2$ .

#### Eksempel 3.3

Vi har målt hastigheden af en faldende sten ved forskellige højder, og finder to datapunkter,  $v_1 = (8 \pm 1)\text{m/s}$  og  $v_2 = (26 \pm 2)\text{m/s}$ , til tiderne  $t_1 = 1.0\text{s}$  og  $t_2 = 3.0\text{s}$ . Her er usikkerheden på hastigheden den ophobede usikkerhed fra både usikkerheden på distance og tiden. Modellen vi vil fitte efter har formen  $v(t) = a \cdot t$ . Det generelle udtryk for vores reducerede chi kvadrat er

$$\chi_{\text{red}}^2 = \frac{1}{N_{\text{dof}}} \left( \frac{(v_1 - v(t_1))^2}{\Delta v_1^2} + \frac{(v_2 - v(t_2))^2}{\Delta v_2^2} \right)$$

Lad os gætte på  $a_1 = 10\text{m/s}^2$ . Altså har vi ikke nogle fit parametre og  $N_{\text{dof}} = N = 2$ . Nu kan vi udregne  $\chi_{\text{red}}^2$ .

<sup>16</sup>Se Barlow kapitel 6.4 for mere om chi-kvadrater.



$$\begin{aligned}\chi_{red}^2 &= \frac{1}{2} \left( \frac{(8\text{m/s} - (10\text{m/s}^2 \cdot 1.0\text{s}))^2}{(1\text{m/s})^2} \right) \\ &\quad + \frac{1}{2} \left( \frac{(26\text{m/s} - (10\text{m/s}^2 \cdot 3.0\text{s}))^2}{(2\text{m/s})^2} \right) \\ &= \frac{1}{2} \left( \frac{2^2}{1} + \frac{4^2}{2^2} \right) = 4\end{aligned}$$

Her får vi et chi kvadrat der er større end 1. Hvis vi i stedet bestemmer accelerationen ved at fitte, så ville vi naturligt finde et endnu bedre chi-kvadrat. For det et fit gør, er netop at minimere chi kvadratet. I det her tilfælde kan man også finde minimummet analytisk, ved at minimere

$$\chi_{red}^2(a) = \frac{1}{2-1} \left( \frac{(v_1 - a \cdot t_1)^2}{\Delta v_1^2} + \frac{(v_2 - a \cdot t_2)^2}{\Delta v_2^2} \right).$$

Hvilket giver  $a = 8.46\text{m/s}^2$ , som præcis er det, som curve\_fit også giver! Dette svarer til et reduceret chi kvadrat på 0.30, hvilket betyder at usikkerheden på  $v$  er overvurderet. Vi kan se, at den resulterende acceleration ligger tættere på den værdi vi ville få ved kun at bruge det andet datapunkt, end hvis vi kun brugte det første datapunkt. Det skyldes at usikkerheden på  $a = v/t$ , der kan regnes ved ophobningsloven, er mindre for  $a = v_2/t_2$  end for  $a = v_1/t_1$ .

Det reducerede chi kvadrat giver kun mening at udregne, når  $N_{dof} \geq 1$ . Ellers giver chi-kvadratet ikke mening at udregne, for hvis man har lige så mange fit parametre som antal datapunkter (et underdetermineret system), så kan man altid analytisk finde et specifikt sæt af parametre, hvor  $\chi^2 = 0$ . Tilsvarende, så hvis vi har flere antal parametre end datapunkter betyde at vi kan lave flere kombinationer af parametrene, der giver  $\chi^2 = 0$ . Men så snart at der er flere punkter end parametre (et overdetermineret system), så er der mange forskellige måder at lave fittet på, og det er højest usandsynligt, at modellen går præcis gennem alle punkter. Generelt så vil  $\chi_{red}^2 > 0$  for et overdetermineret datasæt  $N_{dof} \geq 1$ . Og ofte vil en model give  $\chi_{red}^2 > 1$ , hvis usikkerhederne ikke er overvurderede.<sup>17</sup>

### 3.3 Fitning med $\chi_{red}^2$ og residualplots

I sektionen overfor fandt vi ud af, hvordan man kan give et kvantitativt mål for, hvor godt en model passer til data. Nu vil vi kigge på, hvad man kan gøre, for at forbedre en model. Vi vil gerne finde den mest nøjagtige model til data, fordi det er den, der bedst beskriver fysikken i vores system. En god måde at undersøge, hvordan et fit kan forbedres er at lave et residual plot. Et residual kan findes ved  $y_i - f(x_i)$ , altså data minus model. Residualplottet vil være centreret omkring 0 da chi kvadratet netop er minimeret, og vi ønsker at reducere antallet af punkter, der ligger langt fra 0, ved at se på systematik i residualerne. Ligesom i figur 9, så vil der ofte ligge nogle fysiske betragtninger bag systematikken i residualplotet.

<sup>17</sup>Egentlig bør man ikke forvente at få  $\chi_{red}^2 = 0$  for en perfekt model, hvis selve parametrene i modellen er korrelerede, men det er i de fleste tilfælde stadig en god tommelfingerregel.



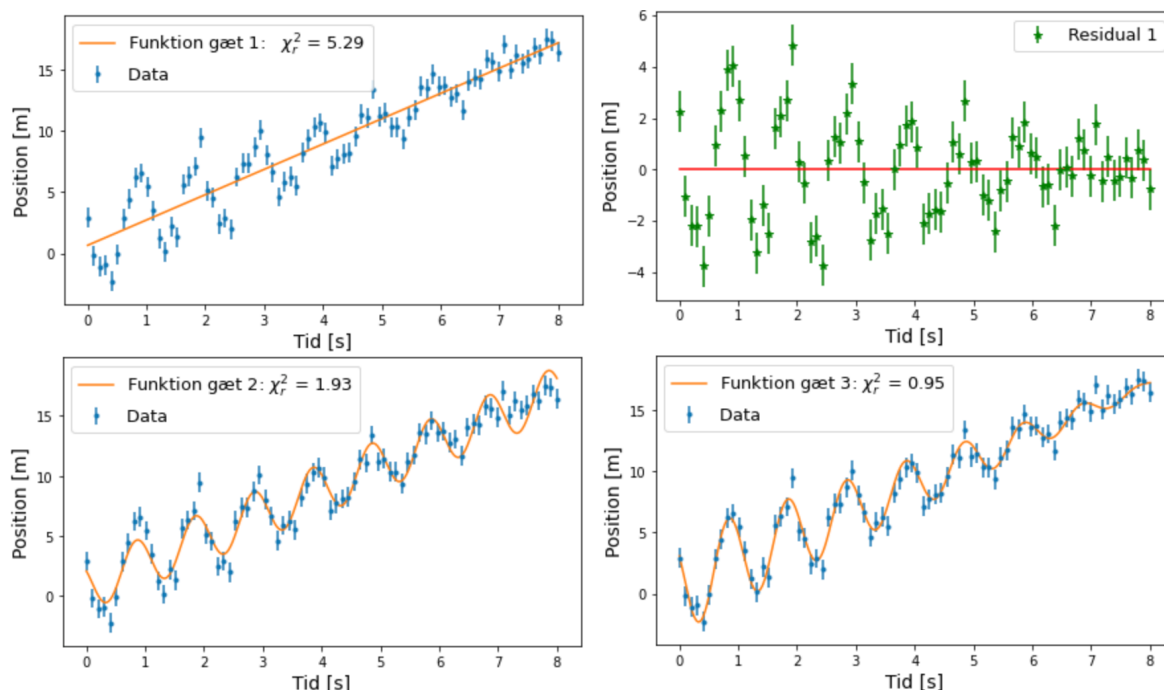


Figure 9: Plots over data og fittet model, samt et residualplot for den første model. Data kommer fra målingen af positionen for et lod ophængt i en fjeder, som bliver trukket opad af en trisse med konstant hastighed. I første omgang gættes på en ret linje. Det giver  $\chi_{red}^2 = 5.29$ . Tydeligvis er modellen forkert, og derfor konstrueres residualplottet (til højre). Fra residualplottet ses at der bør være en sinusfunktion oven i den rette linje. Og faktisk kan man også se at der må være en form for dæmpning, der er nogenlunde lineær. Derfor gætter vi nu på  $x(t) = at + b + A \sin(kt + \phi)$ , hvilket giver  $\chi_{red}^2 = 1.93$ . Derefter gætter vi på  $x(t) = at + b + Ae^{-t/t_0} \sin(kt + \phi)$ , hvilket giver et  $\chi_{red}^2$ , der er helt nede på 0.95.

Er der mange parametre i et fit kan det være svært for *curve\_fit* at finde de bedste parametre. For funktionen starter altid med at gætte på punktet  $\vec{1}$  i parameter rummet, og bevæger sig derfra ved at tage gradienten af  $\chi^2$  i de  $n$  parametre, og finde et minimum. Men der kan være mange lokale minima i det  $n$  dimensionelle parameter rum, og derfor kan man risikere at parametrene bliver helt skæve. Derfor bør man altid plotte den fittede funktion sammen med data. Så er det tydeligt at se, om parametrene er fornuftige. Og ellers kan man give nogle startgæt til *curve\_fit* baseret på aflæsningen af plottet, hvilket gør det mere effektivt at finde det globale minimum.

## 4 Supplerende materiale

Hvis du vil vide mere om databehandling og statistik kan vi anbefale *Statistics - A Guide to the Use of Statistical Methods in the Physical Science* af R. J. Barlow. Bogen bliver også brugt i førsteårskurset Data og Projekt (blok 3), som supplerende materiale i det valgfri kursus Experimental Physics (blok 5) og i kandidatkurset Applied Statistics: From Data to Results (blok 2).

## 5 Formelsamling

1: Den vigtigste formel er udtrykket for usikkerheden på en middelværdi af  $N$  datapunkter.

$$\Delta = \frac{\sigma}{\sqrt{N}}, \quad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

2: Ophobningsloven for den totale usikkerhed på en værdi  $f$  er også meget god.

$$\Delta f = \sqrt{\left(\frac{\partial f}{\partial a} \cdot \Delta a\right)^2 + \left(\frac{\partial f}{\partial b} \cdot \Delta b\right)^2 + \left(\frac{\partial f}{\partial c} \cdot \Delta c\right)^2 \dots}$$

Alle formler for udregning af en samlet usikkerhed på parametre kan udregnes fra ophobningsloven. Her kommer de mest brugte formler.

$$\Delta f = \sqrt{\Delta A^2 + \Delta B^2}, \quad f = f(A, B) = A + B$$

$$\Delta f = a \Delta x, \quad f = f(x) = ax + b$$

$$\Delta f = \frac{df}{dx} \Delta x, \quad f = f(x)$$

3: Når vi vil sammenligne vores eksperiment (A) med et andet (B), eller en tabelværdi, så finder man afstanden mellem de to værdier i enheder af usikkerhederne.

$$z = \frac{D}{\Delta D}, \quad D = |A - B|, \quad \Delta D = \sqrt{\Delta A^2 + \Delta B^2}$$

Hvis  $z$ -værdien, afstanden mellem værdierne i enhed af usikkerheden, er under 1, kan man ofte sige at værdierne stemmer godt overens. For  $z \leq 3$  stemmer de nogenlunde overens og hvis  $z > 3$  er der en signifikant forskel.

4: For at kunne vurdere modeller med  $n$  parametre, der bliver fittet fra  $N$  datapunkter, er det reducerede chi kvadrat meget brugbart.

$$\chi_{red}^2 = \frac{\chi^2}{N_{dof}}, \quad N_{dof} = N - n, \quad \chi^2 = \sum_{i=1}^N \frac{(y_{data,i} - y_{model})^2}{\Delta y_{data,i}^2}$$

Hvis det reducerede chi kvadrat er omkring 1 er modellen god. Får du et reduceret chi kvadrat, der er under 1, så er usikkerhederne nok overvurderede. Er det meget større end 1 skal du granske din data igen og finde en bedre model.

5: En usikkerhed  $\Delta$  og en middelværdi  $\mu$  betyder at der inde for intervallet  $\mu \pm \Delta$  ligger omkring 68% af data, mens der ligger 95% i intervallet  $\mu \pm 2\Delta$  og 99.7% i intervallet  $\mu \pm 3\Delta$ . Plotter man data sammen med middelværdien, så er det derved ikke alle steder at middelværdien vil passe til datapunkterne inde for errorbaren.

Det samme gælder når  $\chi_{red}^2 = 1$ , altså når afvigelsen mellem data og model tilsvare usikkerheden på data. Her vil det kun være omkring 68% af alle datapunkterne, der med usikkerheden ligger inde for funktionsværdien af modellen.