

# GOED METEN MET FOUTEN

EEN INTRODUCTIE IN DE VERWERKING EN  
PRESENTATIE VAN MEETRESULTATEN EN DE  
DISCUSSIE VAN MEETFOUTEN

Herman J.C. Berendsen

Chemische Laboratoria  
(GBB: Moleculaire Dynamica)  
RIJKSUNIVERSITEIT GRONINGEN



# GOED METEN MET FOUTEN

EEN INTRODUCTIE IN DE VERWERKING EN  
PRESENTATIE VAN MEETRESULTATEN EN DE  
DISCUSSIE VAN MEETFOUTEN

Prof. Dr. H. J. C. Berendsen

Chemische Laboratoria  
(GBB: Moleculaire Dynamica)  
RIJKSUNIVERSITEIT GRONINGEN

Berendsen, Herman J. C.

GOED METEN MET FOUTEN, Een introductie in de verwerking en presentatie van meetresultaten en de discussie van meetfouten.

ISBN 90 367 0708 0

Uitg. Bibl. der RU, Groningen NUGI code 113

5e herziene druk, febr. 2009 [1e druk dec. 1997; 2e druk dec. 1999; 3e druk okt. 2000, 4e druk jan 2006]

Copyright ©1997, 1999, 2000, 2006, 2009 H.J.C.B., Chemische Laboratoria, GBB: Moleculaire Dynamica, Rijks Universiteit Groningen.

Correspondentieadres: Nijenborgh 4, 9747 AG, Groningen.

Email: [mdsecr@rug.nl](mailto:mdsecr@rug.nl); [H.J.C.Berendsen@rug.nl](mailto:H.J.C.Berendsen@rug.nl)

Alle rechten voorbehouden. Niets uit deze publicatie mag in welke vorm dan ook worden verveelvoudigd, opgeslagen in een gegevensbestand of openbaar gemaakt, zonder voorafgaande schriftelijke toestemming van de auteur. Het maken van enkelvoudige copieën van gedeelten uit deze uitgave voor persoonlijk gebruik is toegestaan.

Deze uitgave is verkrijgbaar door schriftelijke bestelling bij bovenstaand adres, na overmaking van 12,- euro per exemplaar (inclusief verzendkosten) op ABN/Amro bankrekening 47.45.67.206 van de Faculteit W en N, RuG, onder vermelding "Goed meten met fouten", ref project 801503.

In 2010 zal een vernieuwde Engelse versie verschijnen bij Cambridge University Press [ISBN: 9780521119405 (hardback), 9780521134927 (paperback)]. Zie ook [www.hjcb.nl](http://www.hjcb.nl).



# Inhoudsopgave

<b>voorwoord</b>	<b>vii</b>
<b>samenvatting</b>	<b>1</b>
<b>1 Meetfouten en hun weergave</b>	<b>7</b>
1.1 Inleiding . . . . .	7
1.2 Het weergeven van grootheden met hun foutengrenzen . .	8
1.3 Classificatie van fouten . . . . .	12
1.4 Doorwerking van fouten in de einduitkomst . . . . .	14
<b>2 Waarschijnlijkheidsverdelingen</b>	<b>17</b>
2.1 Eigenschappen van kansverdelingen . . . . .	18
2.2 De binomiaalverdeling . . . . .	20
2.3 De Poissonverdeling . . . . .	21
2.4 De normale verdeling . . . . .	22
2.5 Andere verdelingen . . . . .	24
<b>3 Verwerking van meetseries</b>	<b>27</b>
3.1 De verdelingsfunctie van een serie meetwaarden . . . . .	28
3.2 Het gemiddelde en de gemiddelde kwadratische afwijking van het gemiddelde van een serie meetwaarden . . . . .	30
3.3 Schatting van gemiddelde en variantie . . . . .	33

3.4	De nauwkeurigheid van gemiddelde en variantie en Student's t-verdeling . . . . .	33
3.5	Middelen met ongelijke gewichten . . . . .	37
3.6	Robuuste schattingen . . . . .	40
<b>4</b>	<b>De grafische gegevensverwerking en foutendiscussie</b>	<b>45</b>
4.1	Inleiding . . . . .	45
4.2	Het lineariseren van functies . . . . .	47
4.3	Grafische schattingen van de nauwkeurigheid van parameters in lineaire functies . . . . .	52
4.4	Meetgegevens als ijklijn . . . . .	54
<b>5</b>	<b>Kleinste-kwadratenaanpassing en lineaire regressie</b>	<b>57</b>
5.1	Lineaire regressie . . . . .	57
5.2	De algemene kleinste-kwadratenaanpassing . . . . .	65
5.3	De chi-kwadraattest . . . . .	68
5.4	Nauwkeurigheid van de parameters . . . . .	71
	<b>Bibliografie</b>	<b>77</b>
<b>A</b>	<b>Optellen van fouten</b>	<b>79</b>
<b>B</b>	<b>Systematische fouten in doorwerking</b>	<b>83</b>
<b>C</b>	<b>De binomiaal- en de Poissonverdeling</b>	<b>85</b>
<b>D</b>	<b>Student's t-verdeling</b>	<b>89</b>
<b>E</b>	<b>Schatting van de variantie</b>	<b>91</b>
<b>F</b>	<b>Standaarddeviatie van het gemiddelde</b>	<b>95</b>
<b>G</b>	<b>Weegfactoren bij ongelijke varianties</b>	<b>97</b>

<b>H Kleinste-kwadratenaanpassing</b>	<b>99</b>
<b>Antwoorden</b>	<b>105</b>
<b>Waarschijnlijkheidspapier</b>	<b>109</b>
x: lin 50; y: prob 1-99 . . . . .	110
x: lin 70; y: prob 1-99 . . . . .	111
<b>zakboekblaadjes</b>	<b>113</b>
Chikwadraatverdeling . . . . .	115
Eenheden . . . . .	117
Fysische constanten . . . . .	123
Kleinste kwadratenaanpassing . . . . .	125
Mathematica . . . . .	127
Normale verdeling . . . . .	133
Student's t-verdeling . . . . .	135





# VOORWOORD

*Dit boekje is bedoeld als leidraad voor het presenteren van experimentele gegevens en voor een verantwoorde verwerking van de daarbij optredende meetfouten. Hoewel oorspronkelijk bedoeld voor gebruik bij het (fysisch-chemisch) practicum, is de opzet zo gekozen dat dit boekje ook voor later gebruik bij experimenteel werk - en soms ook bij theoretische simulaties - als naslagwerk kan dienen. Daartoe zijn de belangrijkste aspecten van de foutendiscussie in een samenvatting bijeengebracht en zijn een aantal nuttige gegevens aan het eind toegevoegd. Er wordt ook aandacht besteed aan het juiste gebruik en de juiste notatie van eenheden en fysische constanten.*

*De nadruk ligt op het verwerken en interpreteren van meetgegevens, met een correcte verwerking van meetfouten. Dit boek gaat niet in op het testen van hypothesen of het ontwerpen van experimentele schema's die bedoeld zijn om de invloed van verschillende parameters te testen. Deze onderwerpen komen in alle goede boeken over statistiek aan de orde.*

*De foutendiscussie berust op een aantal regels uit de statistiek, waarvan de statistische achtergrond in de tekst slechts summier wordt toegelicht. Voor degenen bij wie dit een gevoel van onbehagen achterlaat, zijn diepergaande toelichtingen in een aantal bijlagen aan de foutendiscussie toegevoegd. De bijlagen geven niet alleen meer uitleg, maar geven ook meer details, die voor meer geavanceerde toepassingen nuttig kunnen zijn. Er is naar gestreefd om van alle 'moeilijke' regels afleidingen te geven. De bijlagen doen een groter beroep op de wiskundige vaardigheden (bijvoorbeeld lineaire algebra) van de lezer dan de overige tekst. Voor de meeste praktische toepassingen zijn deze bijlagen echter niet nodig.*

*De computer is bij moderne gegevensverwerking en foutendiscussie een nuttig en vaak noodzakelijk hulpmiddel, maar een 'black box' compu-*

terprogramma mag nooit een substituut voor een onbegrepen methode zijn. Er zijn vele programmapakketten die gespecialiseerd zijn in statistische gegevensverwerking. Wij hebben gekozen voor het algemene pakket MATHEMATICA, zodat het 'black box' gedeelte beperkt blijft tot mathematische en grafische functies. Een aantal voorbeeldprogramma's en aanwijzingen voor toepassing zijn aan de tekst toegevoegd.

Dit boekje is mede bedoeld voor zelfstudie en bevat daarvoor een aantal opgaven. Uitwerkingen van de meeste opgaven zijn te vinden achter de bijlagen (pag. 105).

Er zijn ook twee bladen 'waarschijnlijkheidspapier' bijgevoegd, waarop cumulatieve verdelingen kunnen worden uitgezet om te kijken of die aan een normale verdeling voldoen. De grenzen zijn 1% en 99%.

Achterin bevindt zich een aantal 'zakboekblaadjes' met gegevens. Deze zijn als naslag bedoeld om snel een relevant gegeven te kunnen opzoeken. Wanneer u elk blad copiëert onder verkleining tot 71%, d.w.z. van A3 naar A4 formaat, dan ontstaat er een formaat dat past in een standaard ringband van het Succes Junior formaat. (Copyright regels staan het copiëren voor uitsluitend persoonlijk gebruik toe). Plak linkerbladzijden aan de achterkant van de voorgaande rechterbladzijde. Er bestaan perforatoren die zes gaatjes in dat formaat kunnen ponsen.

De tweede druk (2000) bevat een aantal correcties (dank aan Dr A. v.d. Pol) en een uitbreiding met 'robuuste methoden', terwijl de meeste bladen grafiekenpapier uit de eerste druk weggelaten zijn. In de vijfde druk (2009) zijn de fysische constanten geactualiseerd naar de laatste CODATA gegevens (uit 2006) en is de literatuurlijst aangepast.

Het is de bedoeling dat in 2010 een Engelstalige vernieuwde versie zal verschijnen bij Cambridge University Press [ISBN: 9780521119405 (hardback), 9780521134927 (paperback)]. Het oorspronkelijke Nederlandse boekje zal beschikbaar blijven. Zie voor de actuele stand van zaken: [www.hjcb.nl](http://www.hjcb.nl).

De auteur houdt zich ten zeerste aanbevolen voor op- en aanmerkingen van gebruikers van dit boek.

Groningen, februari 2009

H.J.C. Berendsen

# Samenvatting

## Weergeven van meetresultaten met hun nauwkeurigheden

Geef altijd een schatting van de fout in het eindresultaat; geef van het eindresultaat alleen de significante cijfers. Vermeld altijd de eenheid waarin het meetresultaat is uitgedrukt (nooit cursief of vet); gebruik bij voorkeur S.I.-eenheden en gebruik een ondubbelzinnige notatie met de officiële afkortingen (zie blad EENHEDEN, pag. 117). Onderscheid *systematische* en *toevallige* fouten. Geef in een grafiek de foutengrenzen met streepjes aan. De fout  $\Delta x$  in de meetwaarde  $x$  kan als *absolute fout*  $\Delta x$  of als *relatieve fout*  $\Delta x/x$  worden weergegeven. Een fout kan het beste worden gekarakteriseerd met de *standaarddeviatie*  $\sigma$ , die ook wel *middelbare fout* of *r.m.s. (root-mean-square) error* wordt genoemd. Een afwijking van meer dan  $2,5$  à  $3 \times$  de middelbare fout kan als significant (niet toevallig) worden beschouwd.

## Doorwerken van fouten

Onafhankelijke fouten (standaarddeviaties)  $\sigma_x, \sigma_y, \dots$  in meetwaarden  $x, y, \dots$  werken als volgt door in een functie  $f(x, y, \dots)$ :

$$(\sigma_f)^2 = \left(\frac{\partial f}{\partial x}\right)^2 (\sigma_x)^2 + \left(\frac{\partial f}{\partial y}\right)^2 (\sigma_y)^2 + \dots$$

Hieruit volgt dat in sommen en verschillen de *absolute* fouten kwadratisch optellen; in produkten en quotiënten tellen de *relatieve* fouten kwadratisch op. Zijn fouten afhankelijk, dan spelen ook de *covarianties* een rol.

## Verdelingsfuncties en hun eigenschappen

Een meetwaarde  $x_i$  wordt geacht een willekeurige *steekproef* ('sample') uit een waarschijnlijkheidsverdeling  $f(x)$  te zijn. Veel voorkomende

waarschijnlijkheidsverdelingen zijn de *binomiale verdeling* (voor grootheden die twee waarden, zoals *waar* of *onwaar*, kunnen aannemen), de *Poissonverdeling* (voor grootheden die in aantallen worden gemeten) en de *normale verdeling* of Gaussverdeling (voor bijna alle andere gevallen). De eerste twee zijn *discrete* verdelingen, de laatste is een *continue* verdeling, waarbij  $f(x)$  een *waarschijnlijkheidsdichtheid* is (dwz  $f(a) dx$  is de waarschijnlijkheid dat  $x$  in het interval  $(a, a + dx)$  ligt). Waarschijnlijkheidsdichtheden zijn *genormeerd*:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

De *verwachting* of *verwachte waarde* ('expected value') van een grootheid is het gemiddelde van die grootheid over de verdeling. Zo is: het gemiddelde ('mean') van  $x$ :

$$\mu = E[x] = \int_{-\infty}^{\infty} xf(x) dx$$

en de *variantie* ('variance') van  $x$ :

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

$\sigma = \sqrt{\sigma^2}$  is de *standaarddeviatie* ('rms deviation') van de verdeling.  $\mu$  en  $\sigma$  zijn karakteristieken van de verdeling  $f(x)$ , niet van de serie meetwaarden  $x_i$ .

De *cumulatieve verdelingsfunctie*

$$F(x) = \int_{-\infty}^x f(x') dx'$$

geeft de waarschijnlijkheid weer dat de grootheid de waarde  $x$  niet overschrijdt. Deze wordt gebruikt om betrouwbaarheidsintervallen vast te stellen.

### Schattingen uit meetseries

Wordt een meetserie  $x_i, i = 1, \dots, n$  gemeten, waarbij elk meetpunt een onafhankelijke steekproef uit een waarschijnlijkheidsverdeling is, dan kan daaruit een schatting ('estimate') worden verkregen van de karakteristieken van de verdeling. Hiertoe bepaalt men eerst het gemiddelde  $\langle x \rangle$  van de meetreeks:

$$\langle x \rangle = \frac{1}{n} \sum_{i=1}^n x_i$$

en het gemiddelde kwadraat van de afwijking  $\Delta x$  van het gemiddelde:

$$\langle (\Delta x)^2 \rangle = \frac{1}{n} \sum_{i=1}^n (x_i - \langle x \rangle)^2 = \langle x^2 \rangle - \langle x \rangle^2$$

De beste schatting  $\hat{\mu}$  voor het gemiddelde  $\mu$  van de verdeling is nu het gemiddelde van de meetreeks zelf:

$$\hat{\mu} = \langle x \rangle$$

en de beste schatting  $\hat{\sigma}^2$  voor de variantie  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{n}{n-1} \langle (\Delta x)^2 \rangle$$

De beste schatting  $\hat{\sigma}$  voor de standaardafwijking  $\sigma$  van de verdeling is de wortel uit de geschatte variantie.

Het gemiddelde  $\langle x \rangle$  is zelf ook weer een steekproef uit een waarschijnlijkheidsverdeling, die we zouden krijgen als de hele meetserie vaak herhaald zou worden. Deze verdeling heeft een variantie  $\sigma^2/n$  (indien de metingen  $x_i$  statistisch onafhankelijk van elkaar zijn) en de verwachte standaarddeviatie (ook wel *middelbare fout*) in  $\langle x \rangle$  is dus

$$\hat{\sigma}_{\langle x \rangle} = \hat{\sigma} / \sqrt{n}$$

Zijn de metingen niet onafhankelijk, dan heeft dat het effect dat  $n$  effectief kleiner is dan het aantal meetwaarden.

Het betrouwbaarheidsinterval voor het gemiddelde wordt gegeven door *Student's t-verdeling* met  $\nu = n - 1$  vrijheidsgraden.

Wanneer de verschillende meetwaarden geen gelijke gewichten hebben, dan moet bij het bepalen van gemiddelden met deze gewichten  $w_i$  rekening worden gehouden: bijvoorbeeld is het gemiddelde dan

$$\langle x \rangle = \frac{1}{w} \sum_{i=1}^n w_i x_i$$

waar  $w = \sum_{i=1}^n w_i$ . Het gewicht van een meting is omgekeerd evenredig met de variantie, dus met het kwadraat van de standaarddeviatie van die meting.

**Kleinste-kwadratenaanpassing**

Wordt een functie  $f(x)$  aangepast aan een reeks meetwaarden  $(x_i, y_i)$ , dan moeten de parameters in de functie zo worden bepaald dat de (gewogen) som  $S$  van kwadratische afwijkingen tussen meetwaarden  $y_i$  en funktiewaarden  $f_i = f(x_i)$  minimaal is. Hierbij wordt verondersteld dat de fout in  $x_i$  verwaarloosbaar is. De gewichten  $w_i$  moeten omgekeerd evenredig zijn met de varianties  $\sigma_i^2$  van de meetwaarden  $y_i$ .

$$S = \sum_{i=1}^n w_i (y_i - f_i)^2 \text{ minimaal}$$

Zijn alle metingen even nauwkeurig, dan kunnen alle  $w_i$ 's gewoon gelijk aan 1 genomen worden. Aanpassing van een lineaire functie van de parameters aan meetwaarden  $(x_i, y_i)$  heet *lineaire regressie*. We beschouwen aanpassing van de functie  $f(x) = ax + b$ . De parameters  $a$  en  $b$  zijn gegeven door:

$$a = \frac{\langle (\Delta x)(\Delta y) \rangle}{\langle (\Delta x)^2 \rangle}; \quad b = \langle y \rangle - a \langle x \rangle,$$

waar

$$\Delta x = x - \langle x \rangle; \quad \Delta y = y - \langle y \rangle. \quad (1)$$

Zijn de gewichtsfactoren niet gelijk, dan moeten zij in de diverse gemiddelden meegenomen worden.

De schattingen voor de standaarddeviaties (de middelbare fouten) in  $a$  en in  $b$  zijn de wortel uit hun varianties:

$$\hat{\sigma}_a^2 = \frac{S}{w(n-2)\langle (\Delta x)^2 \rangle}$$

$$\hat{\sigma}_b^2 = \hat{\sigma}_a^2 \langle x^2 \rangle$$

Hier is  $w = \sum_{i=1}^n w_i$  het totale gewicht van alle meetpunten samen; zijn alle gewichten  $w_i = 1$ , dan is  $w = n$ . De correlatiecoëfficiënt tussen  $a$  en  $b$  is

$$\rho_{ab} = -\frac{\langle x \rangle}{\sqrt{\langle x^2 \rangle}}$$

Zijn de standaarddeviaties  $\sigma_i$  van alle metingen van te voren goed bekend, dan kan een *chi-kwadraattest* op de som  $\chi^2$  van met  $1/\sigma_i^2$  gewogen afwijkingen worden uitgevoerd:

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - f_i)^2}{\sigma_i^2}$$

Hier is  $\sigma_i^2$  de variantie van  $y_i - f_i$ : als de fout in  $x$  verwaarloosbaar is, is dit de variantie van  $y_i$ ; anders is

$$\sigma_i^2 = \sigma_{y_i}^2 + b^2 \sigma_{x_i}^2$$

$\chi^2$  is dus gelijk aan  $S$  voor het geval dat  $w_i = 1/\sigma_i^2$ . De kleinste-kwadratenanalyse moet dan op deze  $\chi^2$ -som worden uitgevoerd. De gevonden waarde van  $\chi^2$  moet in de buurt van  $n - 2$  liggen; het interval waarover  $\chi^2$  acceptabel is kan uit de tabel op pag. 2 van het blad CHI-KWADRAATVERDELING op pag. 115 worden afgelezen voor diverse acceptatieniveaus.

De *regressiecoëfficiënt*  $r$  die aangeeft hoe goed de punten op een rechte lijn liggen ( $r = \pm 1$ : volledige correlatie;  $r = 0$ : geen correlatie) is

$$r = \frac{\langle(\Delta x)(\Delta y)\rangle}{\sqrt{\langle(\Delta x)^2\rangle\langle(\Delta y)^2\rangle}}$$

Dit is een correlatiecoëfficiënt tussen  $x$  en  $y$ .





# Hoofdstuk 1

## Meetfouten en hun weergave

### 1.1 Inleiding

Het is onmogelijk fysische grootheden zonder fouten te meten. Meestal zijn fouten<sup>1</sup> een gevolg van afwijkingen en onnauwkeurigheden van de meetapparatuur of van de aflezingen van de meetwaarden, maar ook bij ideale apparatuur blijft er een onnauwkeurigheid door ruis (toevallige fluctuaties) in de meetwaarden. Een experimenteel bepaalde grootte heeft dus een bepaalde nauwkeurigheid; het is van belang deze bij het meetresultaat aan te geven omdat bij het gebruik van de gemeten grootte vaak kennis van de nauwkeurigheid nodig is.

Het doel van deze foutendiscussie is om aan te geven hoe men tot een schatting van de fout in het eindresultaat kan komen op grond van de fouten in de meetresultaten waar het eindresultaat van is afgeleid. De grootte van fouten kan vaak uit de statistiek van vele metingen worden afgeleid. In deze korte beschouwing zal niet worden ingegaan op de statistische basis van deze foutenberekening, maar zullen alleen de nodige formules worden gegeven. Wel zijn bijlagen toegevoegd waarin op een aantal details nader wordt ingegaan. De bijlagen gaan uit van een hoger

---

<sup>1</sup>Het woord *fout* (*'error'*) gebruiken we als indifferente indicatie van een afwijking zonder het waardeoordeel dat dit woord als tegenstelling van *goed* inhoudt (zoals het Engelse *'mistake'*).

wiskundig niveau; voor het praktisch gebruik van de foutendiscussie zijn de bijlagen niet nodig.

## 1.2 Het weergeven van grootheden met hun foutengrenzen

Het eindresultaat van een meting moet worden weergegeven met zoveel cijfers als met de nauwkeurigheid overeenkomt. Ook als er nullen aan het eind staan! Dit noemt men de *significante cijfers* van het resultaat. Dat neemt niet weg dat tussenresultaten best (en zelfs beter) met een te hoge precisie kunnen worden gehanteerd om accumulatie van afrondfouten te voorkomen. Geef altijd de nauwkeurigheid in een eindresultaat aan; als die er niet bijstaat wordt stilzwijgend aangenomen dat de fout  $\pm 0,5$  in het laatste cijfer is.

**NB:** In het Nederlands (en in het Duits en Frans) worden decimaalkomma's gebruikt; in het Engels en in alle computernotaties decimaalpunten. Om verwarring te voorkomen wordt het gebruik van punten (Nederlands etc) of komma's (Engels) afgeraden voor het indelen van lange getallen in duizendtallen, zoals 300.000 (Nederlands) of 300,000 (Engels). Gebruik liever een spatie: 300 000.

*Voorbeelden:*

- $1,65 \pm 0,05$
- $2,500 \pm 0,003$
- $35600 \pm 200$ : beter als  $(3,56 \pm 0,02) \times 10^4$
- $5,627 \pm 0,036$  kan ook, maar is alleen zinvol als de fout zelf voldoende nauwkeurig bekend is. Anders geeft men deze waarde aan met  $5,63 \pm 0,04$ .
- De waarde van het getal van Avogadro is bekend als  $(6,022\,136\,7 \pm 0,000\,003\,6) \times 10^{23} \text{ mol}^{-1}$ . Men noteert dit ook wel als  $6,022\,1367(36) \times 10^{23} \text{ mol}^{-1}$ .
- 2,5 betekent  $2,50 \pm 0,05$
- 2,50 betekent  $2,500 \pm 0,005$

- Soms gebruikt men een kleine  $\sigma$  om een fout van ca een kwart in het laatste cijfer aan te geven:  $2,35 = 2,35 \pm 0,03$ , maar dit is niet aan te raden.

Wanneer fouten moeten worden afgerond, doe dit dan altijd naar boven. Geeft een statistische berekening bijvoorbeeld een fout van 0,2476, maak daar dan voor de zekerheid 0,3 van, tenzij de statistiek van de metingen zo goed is dat het geven van twee decimalen (0,25) verantwoord is (zie hiervoor de opmerking aan het eind van par. 3.4, pag. 35). Denk er bij het gebruik van zakrekenmachines aan dat deze van fouten geen weet hebben en meestal een onzinnige precisie suggereren!

Er bestaan verschillende soorten fouten en het moet bij verslaglegging of publikatie duidelijk zijn welk type fout wordt bedoeld. Zonder verdere vermelding wordt altijd de *middelbare fout* of *standaarddeviatie* bedoeld (zie par. 1.3).

Fysische grootheden hebben niet alleen een getalswaarde met bijbehorende precisie, maar ook een *eenheid*. Vermeld altijd de bijbehorende eenheid in de juiste notatie. Over het gebruik van eenheden en de notatie van grootheden zijn internationale afspraken gemaakt. Het afgesproken eenhedenstelsel is het *Système Internationale (S.I.)*.<sup>2</sup> De S.I. eenheden berusten op de grondeenheden m, kg, s, A, K, mol, cd (zie het blad EENHEDEN). Maak er een gewoonte van je strikt aan deze eenheden te houden, ook al wordt er in de internationale (vooral Amerikaanse) literatuur nogal eens van afgeweken. Dus kJ/mol en geen kcal/mol en nm (of pm) en geen Å. Wel is in de chemie de eenheid dm<sup>3</sup> voor volume gebruikelijk om concentraties aan te geven: de molaire (M) is een mol/dm<sup>3</sup> terwijl de S.I. eenheid van concentratie een mol/m<sup>3</sup> is. Gebruik alleen de officiële notaties; dus s en niet sec, g en niet gr of gram, mol en niet mole (*mole* is het Engelse woord voor mol). Gebruik de voorvoegsels voor machten van 10 (micro, Mega, etc., zie het blad EENHEDEN) bij voorkeur voor machten die veelvoudig zijn van 3. In plaats van de hectopascal die in weerberichten wordt gebezigd voor luchtdruk, zou beter de kilopascal (kPa) gebruikt kunnen worden (100 kPa i.p.v. 1000 hectopascal); de reden dat men hectopascals gebruikt is dat de getalswaarden dan hetzelfde zijn als bij de vroeger gebruikelijke en ingeburgerde eenheid bar. Een micron is een onjuiste lengtemaat: gebruik de micrometer ( $\mu\text{m}$ ).

<sup>2</sup>Er is een *Commission for Symbols, Units, Nomenclature, Atomic Masses and Fundamental Constants* (SUN/AMCO) van de International Union of Pure and Applied Physics (IUPAP) die afspraken (ook over fysische constanten) vastlegt en publiceert. De laatste revisie is van 2006 (<http://www.physics.nist.gov/cuu/Constants/>)

Gebruik nooit twee deelstrepen in samengestelde eenheden en vermijdt dubbelzinnige notaties. Dus niet: kg/m/s of kg/m s, maar  $\text{kg m}^{-1} \text{s}^{-1}$ .

Er zijn ook typografische conventies, waarmee je te maken hebt als je werk in druk verschijnt of met een tekstverwerker wordt geproduceerd. Zo worden symbolen voor grootheden *cursief* gedrukt, vectoren *vet* cursief en tensoren met een *schreefloze* letter. Eenheden worden echter altijd met een gewone letter (rechttop) weergegeven, waarbij het gebruik van hoofd- en kleine letters volgens de vastgestelde lijst moet zijn (zie het blad EENHEDEN)

Vaak worden experimentele resultaten weergegeven in de vorm een grafiek. De (middelbare) fout in de meetwaarden kan dan worden aangegeven met een streepje (de lengte van het streepje is twee maal de middelbare fout). Dit kan zowel voor de  $x$ - als voor de  $y$ -waarde worden gedaan, al komt het vaak voor dat één van de waarden zo nauwkeurig is dat een streepje niet zinvol is. Figuur 1.1 geeft een voorbeeld van zo'n grafische presentatie. Hierbij zijn de meetgegevens van tabel 1.1 gebruikt en zijn de gegevens langs de ordinaat op een logaritmische schaal uitgezet, omdat je zo goed kunt zien of er een (verwacht) lineair verband is tussen  $\log c$  en  $t$ . Langs de assen staan getallen; die zijn zelf dimensieloos. Het is een goede gewoonte om bij de assen aan te geven op welke dimensieloze grootheid die getallen slaan: geef dit aan met het symbool (cursief) gedeeld door de eenheid (rechttop), bijvoorbeeld:

$$c/\text{mM of } t/\text{s of } E_{pot}/\text{kJ mol}^{-1}$$

Vermijd meer dan één deelstreep, dus niet  $E_{pot}/\text{kJ/mol}$ . Verdeel de assen zo dat het beschikbare papier zo goed mogelijk benut wordt, maar zorg er tegelijk voor dat de waarden gemakkelijk af te lezen zijn (bijv.: gebruik een cm voor 1, 2 of 5, maar niet voor 3 eenheden).

Er zijn natuurlijk allerlei computerprogramma's om grafieken te tekenen, maar soms gaat een schetsje met de hand op millimeterpapier wel zo snel. Er bestaan verschillende soorten grafiekenpapier met allerlei schaalindelingen. Logaritmische schalen (zoals toegepast in fig. 1.1) kunnen handig zijn, maar met een calculator kan ook gemakkelijk de logaritme op een lineaire schaal worden geplot.

**OPGAVE 1.1** Corrigeer de volgende notaties:

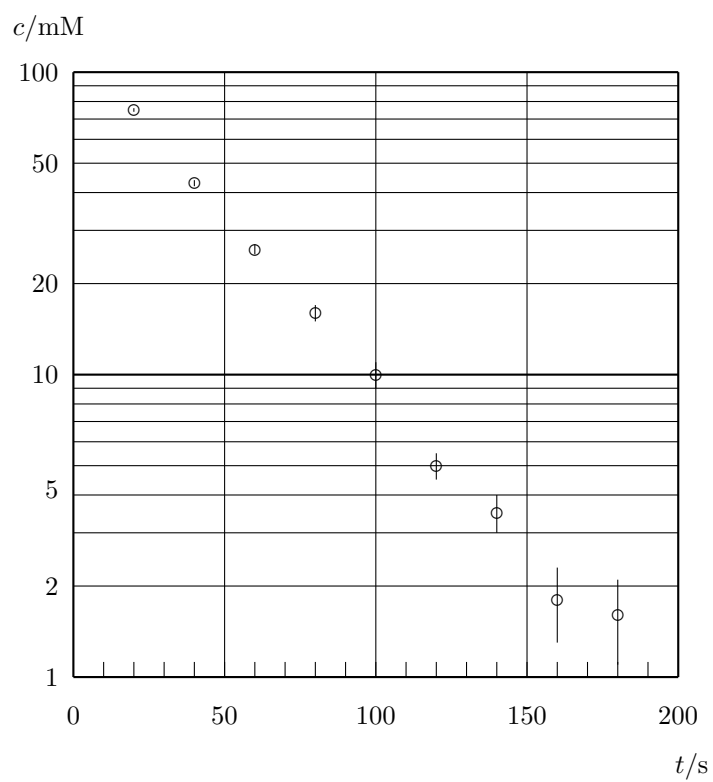
$$l = 3128 \pm 20 \text{ cm}$$

$$c = 0,01532 \text{ M} \pm 0,1 \text{ mM}$$

$$\kappa = 2,52 \times 10^2 \text{ Am}^{-2}/\text{Vm}^{-1}$$

$$k/\text{M}^{-1}/\text{s} = 3571 \pm 2\%$$

$$g = 2 \pm 0,03$$



Figuur 1.1: Concentratie van de reactant uit tabel 1.1 als functie van de tijd uitgezet op een logaritmische schaal. De fout in elke concentratiemeting is gegeven; de fout in de tijdmeting is te klein om weer te geven.

tijd $t/s$	conc. $c/\text{mM} \pm \text{s.d.}$
20	$75 \pm 1$
40	$43 \pm 1$
60	$26 \pm 1$
80	$16 \pm 1$
100	$10 \pm 1$
120	$5.0 \pm 0.5$
140	$3.5 \pm 0.5$
160	$1.8 \pm 0.5$
180	$1.6 \pm 0.5$

Tabel 1.1: Concentratie van een reactant als functie van de tijd. De nauwkeurigheid is weergegeven als geschatte standaarddeviatie.

(voor antwoorden zie blz 105)

### 1.3 Classificatie van fouten

Afgezien van (domme) vergissingen bestaan er systematische en toevallige fouten.

*Systematische fouten* worden bijvoorbeeld veroorzaakt door verkeerde ijkingen van meetinstrumenten, door gebrekkige metingen (parallax bij aflezingen, reactietijd bij tijdmetingen, niet gecorrigeerde nulpuntsfouten), door verontreinigingen in de gebruikte stoffen of door oorzaken die men niet kent. Het laatste komt dan mogelijk aan het licht door resultaten te vergelijken met die van andere onderzoekers in andere laboratoria, die andere apparatuur en liefst ook andere methoden hebben gebruikt. Voor werkelijk kritische experimenten (die bijvoorbeeld een gangbare theorie omverwerpen) is dan ook een onafhankelijke bevestiging vereist.

*Toevallige fouten* zijn uiteraard onvoorspelbaar. Zij vinden hun oorzaak in fysische ruis of fluctuaties door andere oorzaken, en in afleesonauwkeurigheden. Als een meting vele malen herhaald wordt, dan zullen de resultaten een zekere spreiding vertonen rond een gemiddelde. Hieruit kan de geschatte fout in het gemiddelde bepaald worden. De waarschijnlijkheidsverdeling van deze meetwaarden volgt statistische wetmatigheden, waaruit weer regels voor de verwerking van fouten volgen. Heeft

men slechts één of enkele metingen, dan moet de fout geschat worden. Zo kan een op een lineaal afgelezen lengte op  $\pm 0,2$  mm worden geschat, een schuifmaat met nonius kan op  $\pm 0,05$  mm nauwkeurig worden afgelezen en een buret leest men af met een nauwkeurigheid van  $\pm 0,3$  schaaldelen. Bij instrumenten met een digitale uitlezing moet men zich goed bewust zijn van de precisie, omdat de uitlezing zelf meestal te veel decimalen presenteert. De nauwkeurigheid van goede commerciële meetinstrumenten wordt meestal - en soms als een individueel ijkrapport - door de fabrikant aangegeven; in het algemeen wordt de maximale afwijking aangegeven, die een systematisch karakter kan hebben en groter is dan de middelbare toevallige fout.

Het is zaak een gevoel te krijgen voor de grootte van de fouten die tijdens een experiment gemaakt worden. Dan kan men aandacht besteden aan onderdelen die het meest kritisch zijn. Is bij een weegtitratie de druppelfout (t.g.v. de laatste druppel die omslag van de indicator geeft) 10 mg op 5 g of 0,2 %, dan kan men er mee volstaan de titreerspuiten op een 3-decimalenbalans te wegen, want dat geeft al een precisie (bij een weging vóór en één na de titratie van ca 5 gram) van 2 op 5000, of 0,04 %. Beter wegen is onnodige perfectie! Overigens heeft de druppelfout gedeeltelijk een systematisch karakter omdat na de kleuromslag het equivalentiepunt altijd gepasseerd is.

Fouten kan men aangeven als absolute of als relatieve fout. *Absolute fouten* staan in dezelfde dimensie als de gemeten grootte, bijv. een lengte bedraagt  $21,3 \pm 0,2$  mm. *Relatieve fouten* worden gegeven als fractie van de gemeten grootte, of in procenten. Een relatieve fout is een dimensieloos getal. Bijvoorbeeld  $21,3 \text{ mm} \pm 1\%$ . Toevallige fouten  $\epsilon$  zijn *steekproeven* ('samples') uit een waarschijnlijkheidsverdeling en moeten met één of ander gemiddelde worden weergegeven. Men kan de gemiddelde absolute fout  $\langle |\epsilon| \rangle$  vermelden, of de *standaarddeviatie*  $\sigma$ , die ook wel *middelbare* of *r.m.s.* (*root mean square*) fout  $\sqrt{\langle \epsilon^2 \rangle}$  wordt genoemd. De standaarddeviatie (s.d.) is nuttiger omdat daarvoor allerlei statistische betrekkingen gelden.

Bij meer zorgvuldige foutenanalyses kan ook een *betrouwbaarheidsinterval* ('confidence interval') worden gegeven, bijv. een ondergrens en een bovengrens, zodat de waarschijnlijkheid dat de echte waarde tussen die grenzen valt, gelijk is aan een bepaalde fractie, die het *betrouwbaarheidsniveau* ('confidence level') wordt genoemd. Dat kan bijvoorbeeld 50% zijn, of 68% (tussen het gemiddelde  $\pm$  twee standaarddeviaties bij een normale verdeling, zie p.22), of 90% of 99%, naar keuze. Uiteraard



moet bij het vermelden van dergelijke foutengrenzen ook het gebruikte betrouwbaarheidsniveau en de toegepaste methode worden gegeven. Bij voorkeur vermeldt men dan ook het aantal onafhankelijke metingen waaruit het betrouwbaarheidsinterval is afgeleid. Bijvoorbeeld:

$$c = (3,20, 3,35) \text{ mM}; \text{ CONF} = 95\%, n = 5.$$

## 1.4 Doorwerking van fouten in de einduitkomst

Meestal is men geïnteresseerd in een einduitkomst die een functie is van één of meer meetgrootheden. We moeten dan na kunnen gaan hoe de fouten in de meetgrootheden doorwerken in het eindresultaat. Stel dat een evenwichtsconstante  $K$  van een reactie is gemeten als  $K = 305 \pm 5$ . Hoe groot is dan de fout in de standaard vrije enthalpie van de reactie (bij  $T = 300 \text{ K}$ )  $G = -RT \ln K = 14,268 \text{ kJ/mol}$ ? Dit vinden we door differentiatie:

*Is de fout in  $x$  gelijk aan  $\sigma_x$ , dan is de fout  $\sigma_f$  in  $f(x)$  gelijk aan*

$$\sigma_f = \left| \frac{df}{dx} \right| \sigma_x \quad (1.1)$$

In dit voorbeeld is de fout in  $G$  dus gelijk aan  $(RT/K)\sigma_K = 41 \text{ J/mol}$ ; de uitkomst schrijven we als  $G = 14,27 \pm 0,04 \text{ kJ/mol}$ .

Is een fout *samengesteld* uit fouten van meerdere onafhankelijke meetgrootheden, dan moet men de verschillende bijdragen op één of andere manier optellen. Maar gewoon optellen is niet goed: fouten van verschillende onafhankelijke bronnen kunnen + of - zijn en zullen elkaar vaak gedeeltelijk compenseren. De juiste manier van samentellen is het nemen van de wortel uit de som van de kwadraten van de afzonderlijke fouten:

$$\text{Als } f = x + y, \text{ dan is } \sigma_f^2 = \sigma_x^2 + \sigma_y^2 \quad (1.2)$$

of: *onafhankelijke fouten tellen kwadratisch op*. Waarom dit zo is wordt in Bijlage A nader uitgelegd. In het algemeen geldt, wanneer  $f$  een functie is van  $x, y, z, \dots$

$$\sigma_f^2 = \left( \frac{\partial f}{\partial x} \right)^2 \sigma_x^2 + \left( \frac{\partial f}{\partial y} \right)^2 \sigma_y^2 + \dots \quad (1.3)$$

Wanneer de fouten niet onafhankelijk van elkaar zijn, dan spelen ook de covarianties tussen  $x$  en  $y$  een rol (zie voor details Bijlage A):

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} \text{cov}(x, y) + \dots \quad (1.4)$$

Uit vgl 1.3 volgt direct (ga dat na) dat voor optellingen en aftrekkingen de *absolute* fouten (kwadratisch) optellen en dat voor vermenigvuldigen en delingen de *relatieve* fouten (kwadratisch) optellen. Voorbeelden van formule 1.3 zijn:

$f = x + y$ of $f = x - y$	$\sigma_f^2 = \sigma_x^2 + \sigma_y^2$
$f = xy$ of $f = x/y$	$(\sigma_f/f)^2 = (\sigma_x/x)^2 + (\sigma_y/y)^2$
$f = xy^n$ of $f = x/y^n$	$(\sigma_f/f)^2 = (\sigma_x/x)^2 + n^2(\sigma_y/y)^2$
$f = \ln x$	$\sigma_f = \sigma_x/x$
$f = e^x$	$\sigma_f = f\sigma_x$

Wanneer de functie  $f(x)$  een sterke kromming heeft over het gebied waarin de spreiding van  $x$  voorkomt, dan treedt er ook nog een *systematische* afwijking in  $f$  op: de verwachte waarde  $E[f(x)]$  is niet gelijk aan  $f(E[x])$ . Het is een effect dat in de praktijk niet vaak belangrijk is. Bijlage B gaat hier verder op in.

**OPGAVE 1.2** Voer de volgende bewerkingen uit op grootheden met hun standaarddeviaties. De standaarddeviaties van de grootheden zijn met  $\pm$  aangegeven; zij zijn onafhankelijk van elkaar.

a.  $15,000/(5,0 \pm 0,1)$

b.  $(30,0 \pm 0,9)/(5,0 \pm 0,2)$

c.  $\log_{10}(1000 \pm 2)$

d.  $(20,0 \pm 0,3) \exp[-(2,00 \pm 0,01)]$

(voor antwoorden zie blz 105)

**OPGAVE 1.3** Van een eerste-orderreactie wordt de halfwaardetijd bepaald bij vier temperaturen. De temperatuur is nauwkeurig; van de halfwaardetijden zijn de middelbare fouten gegeven:

Temperatuur (graad Celsius)	halfwaardetijd seconden
510	$2000 \pm 100$
540	$600 \pm 40$
570	$240 \pm 20$
600	$90 \pm 10$

Bepaal bij elke temperatuur de snelheidsconstante  $k$  met middelbare fout (welke eenheid?) en zet  $\ln k$  uit tegen de reciproke absolute temperatuur. Bereken de fouten in  $\ln k$  en geef deze met streepjes aan in de grafiek.  
(voor antwoorden zie blz 105)

**OPGAVE 1.4** Stel we meten de versnelling van de zwaartekracht  $g$  door de slingertijd  $T$  van een slinger met lengte  $l$  te meten. Hieruit wordt  $g$  bepaald volgens

$$g = 4\pi^2 l / T^2$$

We meten  $T = 2,007 \pm 0,002$  s en  $l = 1,000 \pm 0,002$  m. Bepaal  $g$  en de nauwkeurigheid van het antwoord. (voor antwoorden zie blz 105)

**OPGAVE 1.5** Men bepaalt de activerings-vrije enthalpie  $\Delta G$  uit de meting van een snelheidsconstante  $k$  met de Eyring formule

$$k = (k_B T / h) \exp(-\Delta G / RT)$$

Hier is  $k_B$  de constante van Boltzmann en  $h$  de constante van Planck (zie het blad FYSISCHES CONSTANTEN).

Hoe werkt een fout in de temperatuur door in  $\Delta G$ ?

Als  $\Delta G = 30$  kJ/mol en  $T = 300$  K, hoe groot is dan de fout in  $\Delta G$  ten gevolge van een fout van  $5^\circ\text{C}$  in de temperatuur? (voor antwoorden zie blz 105)

## Hoofdstuk 2

# Waarschijnlijkheidsverdelingen

Elke meting  $x_i$  van een grootheid  $x$  is in feite een *steekproef* ('sample') uit een waarschijnlijkheidsverdeling of kansverdeling  $f(x)$  van  $x$ . Als  $x$  alleen discrete waarden kan aannemen  $x = k, k = 1, \dots, n$  dan vormt  $f(k)$  een *discrete kansverdeling* en geeft  $f(k)$  de kans aan dat de waarde  $k$  voorkomt. Als  $x$  een continue grootheid is dan is  $f(x)$  een continue functie van  $x$  die de *waarschijnlijkheidsdichtheid* wordt genoemd. De betekenis van  $f(x)$  is: *de kans dat een waarneming  $x_i$  in het interval  $(x, x + dx)$  ligt is gelijk aan  $f(x) dx$* . Om de toevallige afwijkingen in meetwaarden te kunnen analyseren, moeten we iets afweten van de onderliggende waarschijnlijkheidsverdeling.

In dit hoofdstuk bekijken we de eigenschappen van enkele veel voorkomende waarschijnlijkheidsverdelingen: de *binomiaalverdeling*, de *Poissonverdeling* en de *normale verdeling*. In het volgende hoofdstuk gaan we dan na hoe we uit een serie metingen de beste schattingen kunnen maken over karakteristieken van de onderliggende waarschijnlijkheidsverdeling. De werkelijke verdeling zal men nooit precies kunnen bepalen omdat daarvoor oneindig veel metingen nodig zijn.

## 2.1 Eigenschappen van kansverdelingen

Kansverdelingen zijn *genormeed*, d.w.z. de som van alle kansen is gelijk aan 1:

$$\sum_{k=1}^n f(k) = 1 \quad (2.1)$$

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (2.2)$$

Voor de continue verdeling  $f(x)$  nemen we hier aan dat het domein van  $x$  van  $-\infty$  tot  $+\infty$  loopt, maar er zijn ook verdelingen die op een ander domein betrekking hebben (bijv. van 0 tot  $\infty$ ). Waarschijnlijkheden zijn nooit negatief:  $f(k) \geq 0$ ;  $f(x) \geq 0$ .

Het *gemiddelde* van een functie  $g(x)$  van  $x$  over de verdeling heet de *verwachte waarde* ('expected value') of *verwachting*  $E[g]$ <sup>1</sup> van  $g(x)$ :

$$E[g] = \sum_{k=1}^n g(k)f(k) \quad (2.3)$$

$$E[g] = \int_{-\infty}^{\infty} g(x)f(x) dx \quad (2.4)$$

Zo is het gemiddelde ('mean') van  $x$  zelf (welke vaak met  $\mu$  wordt aangeduid):

$$\mu = E[k] = \sum_{k=1}^n kf(k) \quad (2.5)$$

$$\mu = E[x] = \int_{-\infty}^{\infty} xf(x) dx \quad (2.6)$$

De *variantie*  $\sigma^2$  van een verdeling is de verwachting van de kwadratische afwijking van het gemiddelde:

$$\sigma^2 = E[(k - \mu)^2] = \sum_{k=1}^n (k - \mu)^2 f(k) \quad (2.7)$$

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (2.8)$$

---

<sup>1</sup>We gebruiken de notatie  $E[\ ]$  om aan te geven dat  $E$  een *functionaal* is, d.w.z. een functie van een functie.

De wortel uit  $\sigma^2$  heet de *standaarddeviatie* (s.d., ‘standard deviation’)  $\sigma$ . Deze wordt ook wel ‘rms (root-mean-square) deviation’ genoemd. Bij foutendiscussies heet de s.d. die in het antwoord te verwachten is, de *standaardfout* of de *middelbare fout* (‘rms error’).

Dit zijn de belangrijkste gemiddelden van waarschijnlijkheidsverdelingen. Soms worden de *momenten* van een verdeling gebruikt. Het  $n^e$  moment  $\mu_n$  is gedefiniëerd als

$$\mu_n = E[x^n] \quad (2.9)$$

en het  $n^e$  *centrale moment* is

$$\mu_n^c = E[(x - \mu)^n] \quad (2.10)$$

Deze momenten worden wel gebruikt om de mate van afwijking van de normale verdeling te karakteriseren. Zo heet het derde centrale moment in eenheden van de derde macht van de s.d. de ‘skewness’, en het vierde centrale moment is gerelateerd aan de ‘kurtosis’. Aangezien de kurtosis van een normale verdeling gelijk is aan 3, is de ‘excess’ gedefiniëerd als de afwijking van de kurtosis van de normale verdeling. In sommige boeken wordt de naam ‘kurtosis’ gebruikt voor wat hier ‘excess’ heet.

$$\text{skewness} = E[(x - \mu)^3 / \sigma^3] \quad (2.11)$$

$$\text{kurtosis} = E[(x - \mu)^4 / \sigma^4] \quad (2.12)$$

$$\text{excess} = \text{kurtosis} - 3. \quad (2.13)$$

*Cumulatieve* verdelingsfuncties  $F(x)$  geven de kans dat een bepaalde waarde  $x$  niet overschreden wordt:

$$F(k) = \sum_{l=1}^k f(l) \quad (2.14)$$

$$F(x) = \int_{-\infty}^x f(x') dx' \quad (2.15)$$

$F$  is een monotoon toenemende functie met een waarde die altijd tussen 0 en 1 ligt. Cumulatieve verdelingsfuncties zijn nodig voor het bepalen van *betrouwbaarheidsintervallen* en *betrouwbaarheidsgrenzen* (‘confidence limits’). Zo is de waarschijnlijkheid 50% dat  $x$  ligt tussen  $x_1 : F(x_1) = 0,25$  en  $x_2 : F(x_2) = 0,75$ . Voor de waarde van  $x$  die met een waarschijnlijkheid van 1% wordt overschreden, geldt  $F(x) = 0,99$ . De waarde van

$x$  waarvoor  $F(x) = 0,5$  is de *mediaan*; bij  $F(x) = 0,25$  of  $0,75$  spreekt men van *kwartielen* en zo zijn er ook decielen en percentielen.  $x$  is de  $q^e$  ‘quantile’ als  $F(x) = q$ .

Statistische tabellen geven naast de waarschijnlijkheidsdichtheden ook de cumulatieve waarschijnlijkheden. Zij zijn o.a. te vinden in Beyer [10], in Abramowitz and Stegun [9], en in het *Handbook of Chemistry and Physics* [11].

#### MATHEMATICA VOORBEELD 2.1

<<Statistics‘ContinuousDistributions‘ (laadt pakket).

PDF[*dist*,*x*] geeft de waarschijnlijkheidsdichtheid (‘probability density function’) van *dist*.

CDF[*dist*,*x*] geeft de cumulatieve verdelingsfunctie (‘cumulative distribution function’) van *dist*.

Quantile[*dist*,*q*] geeft de  $q^e$  quantiel (de waarde van  $x$  waarvoor  $F(x) = q$ ).

Ook: Mean[*dist*], Variance[*dist*]

StandardDeviation[*dist*],

Skewness[*dist*], Kurtosis[*dist*].

*dist* kan o.a. zijn:

BinomialDistribution[*n*,*p*]

PoissonDistribution[ $\mu$ ]

NormalDistribution[ $\mu$ , $\sigma$ ]

LogNormalDistribution[ $\mu$ , $\sigma$ ]

ChiSquareDistribution[*n*]

StudentTDistribution[*n*]

**OPGAVE 2.1** Bepaal met MATHEMATICA de kurtosis van de normale verdeling. Is dit de kurtosis volgens onze definitie of is dit de ‘excess’?

## 2.2 De binomiaalverdeling

Als een meetgrootheid slechts twee waarden kan aannemen (bijvoorbeeld 0 of 1, onwaar of waar) en elke meting heeft een kans  $p$  dat de uitkomst 1 is, dan is de kans  $f(k)$  dat van  $n$  metingen er precies  $k$  de uitkomst 1 hebben, gelijk aan

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (2.16)$$

Hier is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (2.17)$$

de *binomiaalcoëfficiënt* ‘n over k’ die aangeeft op hoeveel manieren  $k$  objecten uit  $n$  objecten kunnen worden gekozen. Het gemiddelde van  $k$  is  $pn$  en de variantie van  $k$  is  $p(1-p)n$ . Bijlage C legt uit waarom.

**OPGAVE 2.2** Bij een loterij valt op 5% van de loten een prijs. Als ik 10 loten koop, wat is dan de kans dat ik geen prijs, 1 prijs, 2 prijzen, ..., heb? Neem aan dat er heel veel loten zijn, zodat de kans op een prijs niet afhangt van hoeveel prijzen ik heb (dit heet ‘loten met teruglegging’) (voor antwoord zie blz. 105).

**OPGAVE 2.3** Stel de kans dat één meting  $x$  een bepaalde waarde  $x_m$  overschrijdt is 1%. Hoe groot is de kans dat *minstens één* meting uit een serie van 20 onafhankelijke metingen, de waarde  $x_m$  overschrijdt? (voor antwoord zie blz. 105).

## 2.3 De Poissonverdeling

Deze verdeling komen we tegen als *aantallen* geteld worden, zoals een aantal objecten (bijv. bacteria) in een klein volume van een homogene suspensie, of een aantal fotonen dat in een gegeven tijdsinterval  $\Delta t$  met een ‘single photon counter’ wordt geteld, of het aantal atomen dat radioactief vervalft in een bepaalde tijd  $t$ .

Als  $\mu$  het aantal gebeurtenissen is dat *gemiddeld* kan worden verwacht, dan is de kans  $f(k)$  dat we  $k$  gebeurtenissen meten volgens de Poissonverdeling:

$$f(k) = \mu^k e^{-\mu} / k! \quad (2.18)$$

De Poissonverdeling is een limietgeval van de binomiaalverdeling en nadert voor grote  $k$  zelf een normale verdeling (zie Bijlage C). De Poissonverdeling is genormeerd. De verwachting en variantie zijn gegeven door:

$$E[k] = \mu \quad (2.19)$$

$$\sigma^2 = E[(k - \mu)^2] = \mu \quad (2.20)$$

De belangrijkste eigenschap van een Poissonverdeling is dat de s.d. gelijk is aan de wortel uit het gemiddelde. Worden bijv. 10 000 lichtkwanta



gemeten in een bepaald tijdsinterval, dan is de s.d. van deze meting gelijk aan 100, dus de meting heeft een nauwkeurigheid van 1%. Is het gemeten aantal  $k$  groot genoeg ( $> 20$ ) dan is de Poissonverdeling met gemiddelde  $k$  vrijwel gelijk aan een normale verdeling met  $\sigma = \sqrt{k}$ .

**OPGAVE 2.4** Een fotocel produceert een elektrische impuls voor elk geabsorbeerd foton, maar produceert ook pulsen als er geen licht op valt. Het aantal pulsen dat in 1 s wordt geteld bedraagt zonder straling 100 en met straling 900. Hoe groot is de relatieve fout in de gemeten stralingsintensiteit? Hoe groot wordt de relatieve fout in de gemeten stralingsintensiteit wanneer de meting (met en zonder licht) 100 keer wordt herhaald? (antwoord op blz 105.)

## 2.4 De normale verdeling

(zie blad NORMALE VERDELING op pag. 133)

Van de verschillende soorten waarschijnlijkheidsverdelingen komt de normale verdeling veruit het vaakste in de praktijk voor. De reden hiervoor is dat toevallige fluctuaties, die het gevolg zijn van een som van vele onafhankelijke toevallige verstoringen, aan deze verdeling voldoen, ongeacht de verdeling waaraan elk van de deelverstoringen zelf voldoet. Dit heet de *centrale limietstelling*. Voor relatief kleine fluctuaties wordt vrijwel altijd een normale verdeling gevonden; grotere verstoringen kunnen ook wel aan andere verdelingen voldoen.

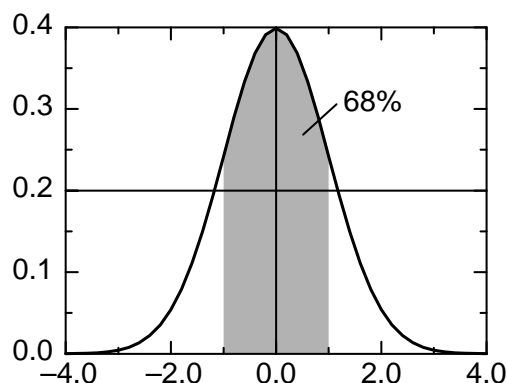
De normale verdeling wordt wiskundig voorgesteld met een zg. *Gaussfunctie*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (2.21)$$

Het gemiddelde is  $\mu$ , de variantie is  $\sigma^2$  en de s.d. is  $\sigma$ .

Figuur 2.1 geeft de normale verdeling weer. De gereduceerde coördinaat  $(x - \mu)/\sigma$  is hier weergegeven: de waarde 0 komt dus overeen met  $x = \mu$  en de waarde 1 met  $x = \mu + \sigma$ . Het grijze oppervlak geeft de waarschijnlijkheid aan dat  $x$  tussen de waarden  $\mu - \sigma$  en  $\mu + \sigma$  valt; dit is gegeven door de cumulatieve verdelingsfunctie  $F(x)$  en is gelijk aan  $F(1) - F(-1) = 1 - 2F(-1) = 0.6826$  (68%).

Tabel 2.1 geeft de kans dat een steekproef  $x$  in een bepaald interval valt en de kans dat  $x$  een bepaalde waarde overschrijdt.



Figuur 2.1: De normale verdeling  $f(x)$  van een grootheid  $x$  met gemiddelde  $\mu$  en s.d.  $\sigma$ . Langs de horizontale as is de grootheid  $(x - \mu)/\sigma$  uitgezet.

We zien dat afwijkingen van meer dan  $2\sigma$  al niet zo vaak voorkomen; afwijkingen van meer dan  $3\sigma$  zijn zeer zeldzaam. Vinden we afwijkingen van minstens  $3\sigma$  dan is het niet waarschijnlijk dat zulke afwijkingen nog toevallig zijn; we noemen de afwijking dan *significant*. Sommigen hanteren  $2,5\sigma$  als de grens voor significante afwijkingen; wat het beste is hangt af van hoeveel metingen men doet<sup>2</sup> (het is helemaal niet significant dat één van de honderd metingen buiten  $2,5\sigma$  valt!), maar het is ook een kwestie van smaak. Het is natuurlijk het beste om het gehanteerde significantieniveau te vermelden.

**OPGAVE 2.5** De centrale limietstelling heeft een handige toepassing: door 12 random getallen  $r$ , die uniform verdeeld zijn over het interval  $(0, 1)$ , bij elkaar op te tellen en van de som 6 af te trekken, krijgen we in goede benadering een steekproef uit een normale verdeling met  $\mu = 0$  en  $\sigma = 1$ :

$$x = \sum_{i=1}^{12} r_i - 6$$

Toon aan dat  $\langle x^2 \rangle = 1$  en genereer met MATHEMATICA een lijst van 100 normaal verdeelde getallen met  $\mu = 0$  en  $\sigma = 1$ .

**OPGAVE 2.6** (Zie het blad NORMALE VERDELING, pag. 133) Wat is de waarschijnlijkheid dat een normaal verdeelde grootheid zich in het interval tussen  $-0,1\sigma$  en  $+0,1\sigma$  t.o.v. het gemiddelde bevindt? (antwoord op pag. 105)

<sup>2</sup>Zie hiervoor de discussie op pag. 41 en de tabel op pag. 2 van het blad NORMALE VERDELING.

afwijking $\Delta$ in eenh. $\sigma$	kans in $(\mu - \Delta, \mu + \Delta)$	kans op $> \mu + \Delta$
0,6745	50 %	25 %
1	68,3 %	15,9 %
1,5	86,6 %	6,68 %
2	95,45 %	2,28 %
2,5	98,76 %	0,62 %
3	99,73 %	0,135 %
4	99,993 66 %	0,003 17 %
5	99,999 943 %	0,000 029 %

Tabel 2.1: Kans dat een steekproef uit een normale verdeling in het interval  $(\mu - \Delta, \mu + \Delta)$  valt, en kans dat een steekproef groter dan  $\mu + \Delta$  (of kleiner dan  $\mu - \Delta$ ) is.

**OPGAVE 2.7** (Zie het blad NORMALE VERDELING, pag. 133)  
Bepaal met een zakrekenmachine de waarde van de Gaussfunctie voor  $x/\sigma = 6$  en de kans dat deze waarde overschreden wordt met de benaderingsformule voor grote  $x$  (vermeld in het blad NORMALE VERDELING). Is deze formule hier geldig? (antwoord op pag. 105)

## 2.5 Andere verdelingen

Er zijn nog vele andere verdelingen. Sommige daarvan komen wij nog tegen: zij zijn belangrijk voor het vaststellen van betrouwbaarheidsintervallen van uit metingen afgeleide grootheden.

- De **Multinomiaalverdeling** geeft de kansverdeling wanneer er een keuze uit een aantal (meer dan twee) elkaar uitsluitende discrete mogelijkheden wordt gemaakt, ieder met een eigen waarschijnlijkheid. Zie Bijlage C, blz. 85.
- De **Chi-kwadratverdeling** geeft de verdeling van de som  $\chi^2$  van het kwadraat van een aantal normaal verdeelde variabelen. Deze verdeling wordt gebruikt om betrouwbaarheidsintervallen te verkrijgen voor de afwijkingen tussen meetwaarden en voorspelde waarden in het geval dat de s.d. van de meetwaarden bekend is. Zie sectie 5.4 op blz 71 en het blad CHI-KWADRAATVERDELING op blz 115.

- **Student's t-verdeling** is de verdeling van het quotiënt van een normaal verdeelde variabele en een  $\chi^2$ -verdeelde variabele. Deze verdeling wordt gebruikt om betrouwbaarheidsintervallen te verkrijgen voor het gemiddelde van een reeks waarnemingen, waarbij de s.d. van de meetwaarden niet van tevoren bekend is. Zie sectie 3.4 op blz 33, Bijlage D op blz 89 en het blad STUDENT'S T-VERDELING op blz 135.
- De **F-verdeling** geeft de verdeling van het quotiënt van twee  $\chi^2$ -verdeelde variabelen. Deze verdeling, die onder meer gebruikt wordt om twee theorieën te vergelijken, gegeven een set data, wordt hier niet verder behandeld.
- De **log-normale verdeling** is een normale verdeling van  $\log x$  in plaats van  $x$ . Deze verdeling is vooral van toepassing op grootheden die nooit negatief kunnen worden, zoals een concentratie, een lengte, of een tijdsinterval.



## Hoofdstuk 3

# Verwerking van meetseries

Stel we hebben een reeks gelijkwaardige waarnemingen  $x_i$ , die alleen van elkaar verschillen door afwijkingen van toevallig karakter. We nemen aan dat de waarschijnlijkheidsverdeling van deze afwijkingen een normale verdeling is, die gekarakteriseerd wordt door een gemiddelde  $\mu$  en een standaarddeviatie  $\sigma$  of variantie  $\sigma^2$ . Hoewel we de echte waarschijnlijkheidsverdeling niet precies kunnen bepalen omdat we maar een beperkt aantal meetwaarden hebben (en deze zijn slechts steekproeven ('samples') uit de verdeling), kunnen we wel de best mogelijke *schatting* ('estimate') maken van  $\mu$  en  $\sigma$ . Deze geschatte waarden worden vaak met een dakje aangegeven:  $\hat{\mu}, \hat{\sigma}$ . We zullen eerst naar de verdeling als geheel kijken (par. 3.1) en daarna aangeven hoe uit eigenschappen van de meetgegevens (par. 3.2) de eigenschappen van de verdelingsfunctie geschat kunnen worden (par. 3.3). Vervolgens gaan we na hoe nauwkeurig de schatting voor het gemiddelde (dus het meetresultaat waarin we geïnteresseerd zijn) eigenlijk is (par. 3.4).

### 3.1 De verdelingsfunctie van een serie meetwaarden

Om een indruk van de verdelingsfunctie te krijgen kunnen we de meetwaarden uitzetten in een *histogram*. Daarvoor sorteren we de meetwaarden eerst in opklimmende volgorde en groeperen ze vervolgens in vooraf vastgestelde intervallen. Het staafjesdiagram van het aantal waarnemingen per interval heet een histogram.

*Voorbeeld.* Stel we herhalen een experiment  $25\times$  en meten de waarden die in tabel 3.1 zijn gegeven.

4,1	3,1	4,9	5,5	6,1
6,2	7,6	6,8	4,7	3,5
2,3	5,0	2,7	8,5	5,4
6,7	4,4	5,8	5,9	4,8
5,1	5,9	4,5	3,8	7,4

Tabel 3.1: Een serie van 25 meetwaarden.

Als we deze waarden in opklimmende grootte sorteren krijgen we de reeks van tabel 3.2. MATHEMATICA kent hiervoor het commando `Sort[list]`.

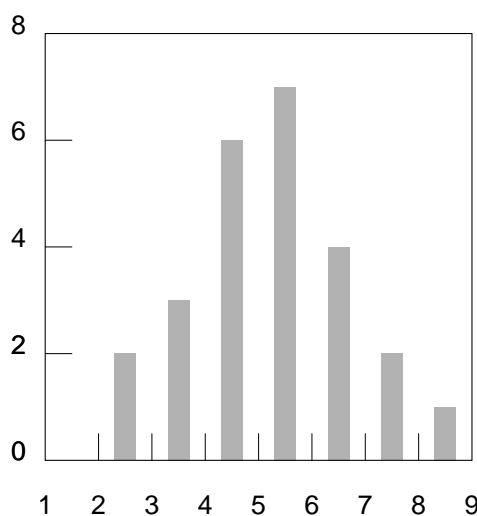
1: 2,3	6: 4,1	11: 4,9	16: 5,8	21: 6,7
2: 2,7	7: 4,4	12: 5,0	17: 5,9	22: 6,8
3: 3,1	8: 4,5	13: 5,1	18: 5,9	23: 7,4
4: 3,5	9: 4,7	14: 5,4	19: 6,1	24: 7,6
5: 3,8	10: 4,8	15: 5,5	20: 6,2	25: 8,5

Tabel 3.2: Een gesorteerde reeks meetwaarden.

Deze kunnen we gemakkelijk in intervallen ('bins') verdelen, bijvoorbeeld  $[2, 3)$ ,  $[3, 4)$ ,  $[4, 5)$ , etc. De gevonden aantallen in elk interval leveren het histogram van fig. 3.1 op.

Zo'n histogram is een benadering van de complete verdeling  $f(x)$ . Het heeft alleen zin een histogram op te stellen wanneer er een groot aantal meetwaarden ter beschikking staan, omdat bij een klein aantal metingen toch geen betrouwbare conclusies over de verdeling te trekken zijn.

De analyse van een verdelingsfunctie kan gemakkelijker gemaakt worden



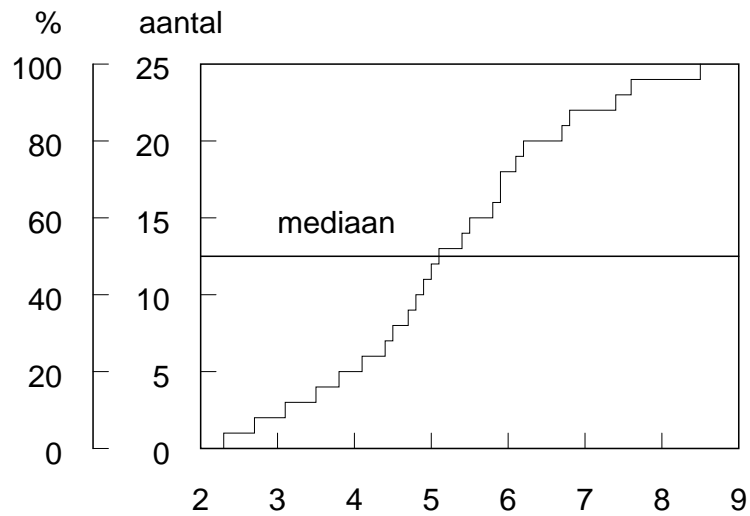
Figuur 3.1: Histogram van de meetgegevens uit de tabel.

via een *cumulatieve verdelingsfunctie*. Deze wordt grafisch weergegeven door het volgnummer in de geordende reeks waarnemingen uit te zetten tegen de waarde van  $x$ . Voor het hierboven gegeven voorbeeld geeft dat fig. 3.2.

De mediaan van de verdeling is de waarde van  $x$  waar evenveel metingen onder als boven liggen (bij volgnummer 13, dus waarde 5,1 in het voorbeeld). Dit hoeft niet hetzelfde te zijn als het gemiddelde. De mediaan is de 50<sup>e</sup> *percentiel* van de verdeling. De  $n^e$  percentiel is de waarde van  $x$  waar  $n\%$  van de metingen onder liggen.

Om een gemeten verdelingsfunctie met een theoretische te vergelijken kan het beste naar de cumulatieve functies worden gekeken. Men kan voor de verticale as van de cumulatieve verdeling een zodanige niet-lineaire schaal kiezen dat de cumulatieve functie voor een bepaalde theoretische verdeling een rechte lijn wordt, zodat afwijkingen gemakkelijk te constateren zijn. Er bestaat zg. *waarschijnlijkheidspapier* met zo'n schaalindeling voor de normale verdeling (zie blz. 110,111). Wordt een experimentele cumulatieve verdelingsfunctie op waarschijnlijkheidspapier uitgezet, zoals voor het hier gegeven voorbeeld is gedaan in fig. 3.3, dan kunnen  $\hat{\mu}$  en  $\hat{\sigma}$  uit de grafiek worden afgelezen.





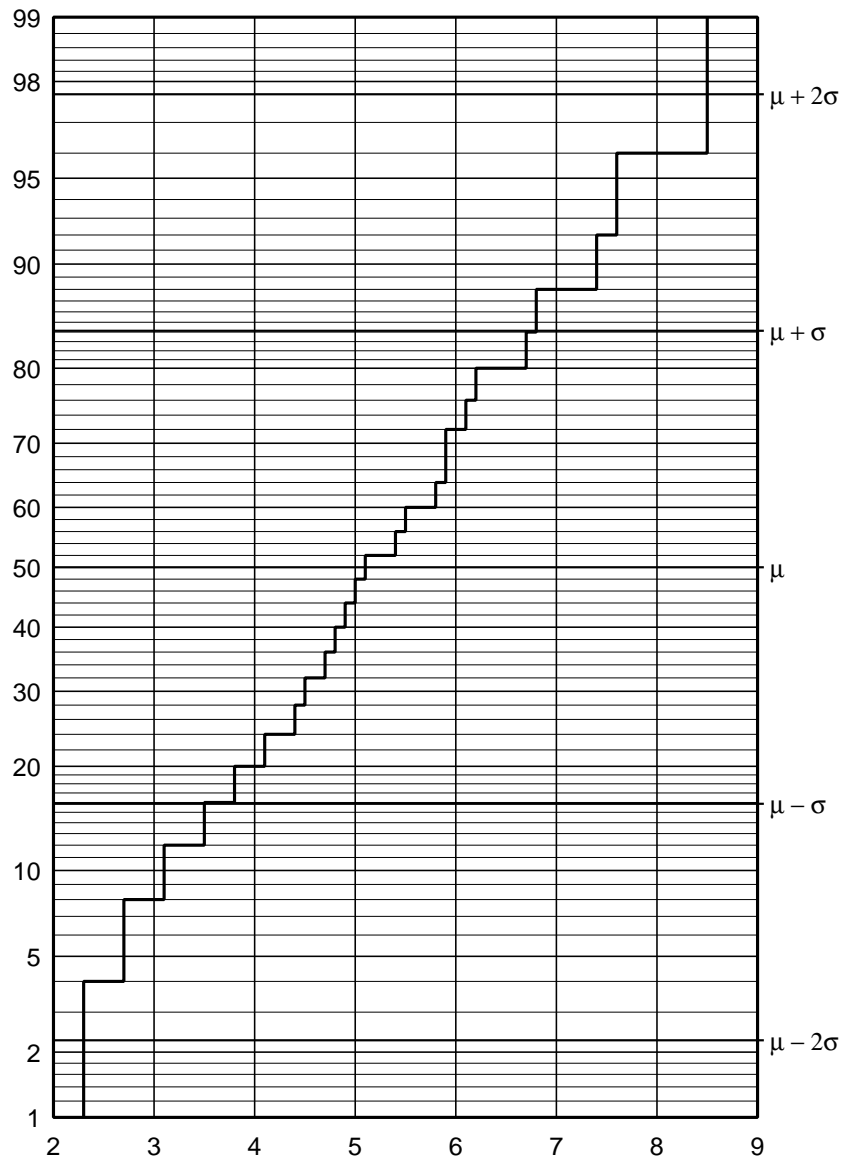
Figuur 3.2: Cumulatieve verdelingsfunctie van de meetgegevens, met mediaan van de verdeling.

**OPGAVE 3.1** Voldoen de gegevens van de tabel aan een normale verdeling? Zo ja, schat dan  $\hat{\mu}$  en  $\hat{\sigma}$  door een rechte lijn door de cumulatieve verdelingsfunctie te trekken.

### 3.2 Het gemiddelde en de gemiddelde kwadratische afwijking van het gemiddelde van een serie meetwaarden

Gemiddelden over een serie meetwaarden geven we hier aan met  $\langle \rangle$  (bijv.  $\langle x \rangle$ ). Vaak worden gemiddelden genoteerd met een streepje boven de variabele, bijv.  $\bar{x}$ . Om eigenschappen van de waarschijnlijkheidsverdeling - waaruit de meetwaarden steekproeven zijn - te schatten, hebben we de volgende gemiddelden nodig:

Het *gemiddelde*  $\langle x \rangle$  van een serie gelijkwaardige, onafhankelijke meet-



Figuur 3.3: Cumulatieve verdeling van de meetgegevens, uitgezet op waarschijnlijkheidspapier.

waarden  $x_i, i = 1, \dots, n$ , is gegeven door

$$\langle x \rangle = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.1)$$

Wanneer we de afwijking van het gemiddelde aanduiden met  $\Delta x_i$ :

$$\Delta x_i = x_i - \langle x \rangle, \quad (3.2)$$

dan is de *gemiddelde kwadratische afwijking van het gemiddelde* gegeven door

$$\langle (\Delta x)^2 \rangle = \frac{1}{n} \sum_{i=1}^n (\Delta x_i)^2 \quad (3.3)$$

Om dit gemiddelde te bepalen, moeten de meetwaarden twee maal doorlopen worden (eerst om  $\langle x \rangle$  te bepalen en daarna om  $\langle (\Delta x)^2 \rangle$  te bepalen). Dit kan voorkomen worden door de volgende formule te gebruiken:

$$\langle (\Delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2, \quad (3.4)$$

waar

$$\langle x^2 \rangle = \frac{1}{n} \sum_{i=1}^n x_i^2 \quad (3.5)$$

*Opmerking:* Als de  $x_i$ 's grote getallen zijn met een relatief kleine spreiding, dan kan formule 3.4 door afrondfouten wel eens numeriek onnauwkeurige antwoorden geven op een computer. Daarom is gebruik van formule 3.4 in het algemeen niet aan te raden. De remedie is van alle  $x$ -waarden een constante die dicht bij  $\langle x \rangle$  ligt, bijvoorbeeld de eerste meetwaarde, af te trekken. Bij berekening van het gemiddelde moet daarvoor natuurlijk later weer gecorrigeerd worden.

**OPGAVE 3.2** Bewijs vergelijking 3.4.

**OPGAVE 3.3** Als van alle  $x$ -waarden eerst een constante wordt afgetrokken, moet daarvoor na berekening van de gemiddelde kwadratische afwijking volgens formule 3.4 dan later gecorrigeerd worden?

### 3.3 Schatting van gemiddelde en variantie

De bovenstaande gemiddelden zijn eenvoudig eigenschappen van de serie meetwaarden. Maar die willen we nu gebruiken om schattingen te maken van het gemiddelde en de variantie (en dus ook de standaarddeviatie) van de waarschijnlijkheidsverdeling waaruit de meetwaarden willekeurige steekproeven zijn.

Voor het gemiddelde is het antwoord eenvoudig: de beste schatting  $\hat{\mu}$  voor het gemiddelde van de verdeling is het gemiddelde  $\langle x \rangle$  van de serie meetwaarden zelf:

$$\hat{\mu} = \langle x \rangle \quad (3.6)$$

Voor de variantie ligt dat iets moeilijker. Stel dat we de variantie niet van te voren kennen en deze moeten afleiden van de spreiding van de meetwaarden zelf. De beste schatting  $\hat{\sigma}^2$  voor de variantie is dan iets groter dan de gemiddelde kwadratische afwijking van het gemiddelde van de meetwaarden:

$$\hat{\sigma}^2 = \frac{n}{n-1} \langle (\Delta x)^2 \rangle \quad (3.7)$$

De reden dat hier  $n/(n-1)$  staat is dat  $\langle x \rangle$  niet gelijk is aan  $\mu$ , maar zelf gecorreleerd is met de waarden uit de meetserie (zie voor de afleiding hiervan Bijlage E). De formule geldt alleen als de meetwaarden onafhankelijk van elkaar zijn (anders is  $\hat{\sigma}^2$  nog groter), maar daar waren we vanuit gegaan. Overigens is de faktor  $n/(n-1)$  niet erg belangrijk, vooral bij grote  $n$ .

De beste schatting voor de standaardafwijking van de verdeling  $\hat{\sigma}$  is de wortel uit  $\hat{\sigma}^2$ :

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \quad (3.8)$$

### 3.4 De nauwkeurigheid van gemiddelde en variantie en Student's t-verdeling

De nauwkeurigheid van het gemiddelde is niet gelijk aan  $\sigma$ , maar volgt daar wel uit. Hoe meer metingen er gedaan zijn, hoe nauwkeuriger het gemiddelde van de meetwaarden gelijk wordt aan de verwachting  $\mu$  van de echte verdelingsfunctie. Het gemiddelde  $\langle x \rangle$  is zelf ook weer een steekproef uit een verdelingsfunctie, die we zouden vinden als we de hele meetserie een groot aantal keren zouden herhalen. Wanneer vele series van  $n$

onafhankelijke metingen zouden zijn verricht, dan zou de variantie van het gemiddelde gelijk zijn aan

$$\sigma_{\langle x \rangle}^2 = \sigma^2/n. \quad (3.9)$$

Zie voor de afleiding hiervan Bijlage F. Dus is de *schatting* voor de standaardafwijking  $\hat{\sigma}_{\langle x \rangle}$  van het gemiddelde  $\langle x \rangle$  (ook wel de *middelbare fout* of '*rms error*' van  $\langle x \rangle$  genoemd):

$$\hat{\sigma}_{\langle x \rangle} = \hat{\sigma}/\sqrt{n}. \quad (3.10)$$

Ook deze formule geldt alleen wanneer de meetwaarden onafhankelijk zijn. Zijn zij afhankelijk dan tellen de afzonderlijke fluctuaties niet meer kwadratisch op en wordt de middelbare fout *groter*. Voor het veel voorkomende geval dat de afhankelijkheid van een serie meetpunten veroorzaakt wordt door correlatie van opeenvolgende meetpunten, kan men een *correlatielengte*  $n_c$  definiëren. De formules blijven dan geldig, maar in plaats van het aantal meetpunten  $n$  moet nu  $n/n_c$  worden ingevuld. Zie voor meer details Bijlage F.

Wanneer de metingen steekproeven uit een normale verdeling zijn, dan zou men wellicht verwachten dat de grootheid

$$t = \frac{\langle x \rangle - \mu}{\hat{\sigma}/\sqrt{n}} \quad (3.11)$$

ook aan een normale verdeling zal voldoen. Dat is echter niet het geval, omdat  $\hat{\sigma}$  niet gelijk is aan de echte  $\sigma$  van de verdeling en zelf ook weer een spreiding kent. Houd men daar rekening mee, dan vindt men dat  $t$  aan een verdeling voldoet die *Student's t-verdeling* wordt genoemd<sup>1</sup> en die op het blad STUDENT'S T-VERDELING (135) is weergegeven (voor een afleiding zie Bijlage D op blz 89). In de limiet van grote aantallen meetpunten wordt de t-verdeling gelijk aan de normale verdeling, maar voor kleine aantallen is de t-verdeling breder. De t-verdeling heeft het aantal *vrijheidsgraden*  $\nu$  als parameter; dit is één minder dan het aantal meetpunten:

$$\nu = n - 1. \quad (3.12)$$

Men zou kunnen zeggen dat één meetpunt al 'verbruikt' is om het gemiddelde te bepalen (net als bij de schatting van  $\sigma$ , zie formule 3.7). Om

---

<sup>1</sup>W.S. Gosset, *The probable error of a mean*, Biometrika **6** (1908) 1. 'Student' was een bijnaam van de Engelse statisticus W.S.Gosset (geb. 1876).

zonder verdere informatie iets over de nauwkeurigheid van een gemiddelde te kunnen zeggen moet men natuurlijk minstens twee meetpunten hebben.

Bij gebruik van de t-verdeling kan men het beste een *betrouwbaarheidsinterval* geven, bijv. de grenzen waartussen het echte gemiddelde verwacht wordt te vallen met een waarschijnlijkheid van 50% (of 80%, 90%, 95%, 99%, ..., naar keuze).

Tot slot geven we nog een indicatie voor de betrouwbaarheid van  $\hat{\sigma}$ : als de metingen onafhankelijk zijn en de afwijkingen zijn steekproeven uit een normale verdeling, dan is de *relatieve* nauwkeurigheid van  $\hat{\sigma}$  gelijk aan  $1/\sqrt{2(n-1)}$ . Bijlage F geeft meer details. Hetzelfde geldt dan ook voor de relatieve nauwkeurigheid van de berekende middelbare fout in het gemiddelde. Vind men bijvoorbeeld als uitkomst van een serie van 10 onafhankelijke metingen  $5,367 \pm 0,253$  dan dient men dit op te geven als  $5,4 \pm 0,3$  omdat de relatieve nauwkeurigheid van het getal 0,253 dan  $1/\sqrt{18}$  of 24% (= 0,06) bedraagt (onvoldoende nauwkeurig voor twee decimalen). Was deze uitkomst echter het resultaat van 100 onafhankelijke metingen, dan kan men als antwoord  $5,37 \pm 0,25$  geven, omdat de relatieve nauwkeurigheid van het getal 0,253 dan  $1/\sqrt{198}$  of 7% (= 0,02) bedraagt (voldoende nauwkeurig voor twee decimalen). Tabel 3.3 geeft de relatieve s.d. in procenten van  $\hat{\sigma}$  voor verschillende waarden van het aantal onafhankelijke metingen  $n$ . Dezelfde relatieve nauwkeurigheid geldt ook voor de s.d. van het gemiddelde, zoals die met formule 3.10 wordt berekend.

$n$	s.d. ( $\hat{\sigma}$ ) %	$n$	s.d. ( $\hat{\sigma}$ ) %	$n$	s.d. ( $\hat{\sigma}$ ) %
2	70	10	24	50	10.1
3	50	15	19	60	9.2
4	41	20	16	70	8.5
5	35	25	14	80	8.0
6	32	30	13	90	7.5
7	29	35	12	100	7.1
8	27	40	11	150	5.8
9	25	45	11	200	5.0

Tabel 3.3: *Relatieve nauwkeurigheid (s.d.) van de geschatte standaarddeviatie van een verdeling op grond van een reeks van  $n$  onafhankelijke metingen.*

Is de standaarddeviatie al niet erg nauwkeurig vast te stellen, een schatting van *skewness* of *excess* is vaak nauwelijks zinvol. Deze schattingen met hun s.d. zijn

$$\text{skewness} = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i}{\hat{\sigma}} \right)^3 \pm \sqrt{\frac{15}{n}} \quad (3.13)$$

$$\text{excess} = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i}{\hat{\sigma}} \right)^4 - 3 \pm \sqrt{\frac{96}{n}}. \quad (3.14)$$

**OPGAVE 3.4** *Maak deze opgave eerst op papier. Controleer vervolgens de antwoorden met het MATHEMATICA programma analyse dat hieronder wordt gegeven (1).*

Bereken met de getallen uit het voorbeeld van par. 3.1 (tabel 3.1) achtereenvolgens:

1. het gemiddelde,
2. de gemiddelde kwadratische afwijking van het gemiddelde,
3. de beste schatting van de variantie van de verdeling,
4. de beste schatting van de standaarddeviatie van de verdeling,
5. de beste schatting voor de middelbare fout (s.d) in het gemiddelde,
6. de nauwkeurigheid van die middelbare fout.

Vergelijk het gevonden gemiddelde en de gevonden standaardafwijking met de resultaten van opgave 3.1. Geef nu het gemiddelde met zijn nauwkeurigheid in het juiste aantal cijfers. Geef ook de 50% en 90% betrouwbaarheidsintervallen voor het gemiddelde op grond van de Student's t-verdeling. Voor antwoord zie blz 105.

#### MATHEMATICA VOORBEELD 3.1

Dit programma **analyse** berekent van een lijst getallen  $x$  de beste schattingen voor het gemiddelde en de variantie en standaarddeviatie, elk met hun nauwkeurigheden. Als er voldoende punten zijn worden ook de 'skewness' en de 'excess' uitgerekend. Er wordt een plotje van de punten gemaakt.

```
Off[General::spell1]; ClearAll[analyse];
analyse[x_] := Module[
  {dxsqav, exc, n,result,sd,sdav, skew, xa, xav, xx, var},
```

```

n=Length[x];
If[n < 3,(Print[""]; Print["Minder dan 3 punten"]; Return[])];
result = Table[,{4}];
xav = Apply[Plus,x]/n; xa = x-xav; xx=xa xa;
dxsqav= Apply[Plus,xx]/n; var = n/(n-1) dxsqav;
sd=Sqrt[var]; sdav=Sqrt[var/n];
result[[1]] = xav; result[[2]] = sdav; result[[3]] = sd;
  result[[4]] =var;
Print[""]; Print["Resultaat van analyse:"];
Print["Ongewogen middeling van ",N[n,4],"punten"];
Print["  gemiddelde = ",N[xav,6]," +/- ",N[sdav,6]];
Print["  stand.dev. = ",N[sd,6]," +/- ",
  N[sd/Sqrt[2. (n-1)],6]];
Print["  variantie = ",N[var,6]," +/- ",
  N[var Sqrt[2/(n-1)],6]];
If[n>=20,
  (skew=Apply[Plus,xx xa]/(n var sd);
  Print["  skewness = ",N[skew,3]," +/- ",N[Sqrt[15/n],3]],
  Print["  skewness: onvoldoende statistiek"]);
If[n>=100,
  (exc=(xx.xx)/(n var var)-3.;
  Print["  excess = ",N[exc,2]," +/- ",N[Sqrt[96/n],2]],
  Print["  excess: onvoldoende statistiek"]);
Print[""]; Print["Plot: punten x-xav, met 1x en 2x s.d." ];
points = Graphics[Table[Line[{{i,xa[[i]},{i+1,xa[[i+1]]}},
  {i,n-1}]];
lines = Graphics[{{Line[{{0,0},{n,0}},Line[{{0,sd},{n,sd}},
  Line[{{0,-sd},{n,-sd}},Line[{{0,2*sd},{n,2*sd}},
  Line[{{0,-2*sd},{n,-2*sd}}]}];
Show[points,lines];
Return[result];
]

```

### 3.5 Middelen met ongelijke gewichten

Tot nu toe gingen we er van uit dat alle meetwaarden hetzelfde gewicht hebben, d.w.z. steekproeven uit dezelfde verdelingsfunctie zijn. Maar er kunnen zich gevallen voordoen dat de ene meetwaarde nauwkeuriger is dan de andere; in dat geval moet de nauwkeurigste waarde een groter gewicht in de middeling krijgen dan de minder nauwkeurige. Dit kan zich voordoen wanneer dezelfde grootheid op verschillende manieren is



gemeten en achteraf het beste gemiddelde moet worden bepaald. Elk van de meetwaarden kan dan een andere standaarddeviatie (of middelbare fout) hebben.

De gewichtsfactor  $w_i$  is omgekeerd evenredig met het kwadraat van de standaarddeviatie van de meetwaarde (zie voor een verklaring hiervan Bijlage G). Gewogen middeling betekent

$$\langle x \rangle = \frac{1}{w} \sum_{i=1}^n w_i x_i; \quad w = \sum_{i=1}^n w_i. \quad (3.15)$$

Dit geldt niet alleen voor  $x$  maar voor elke te middelen grootheid, bijvoorbeeld:

$$\langle x^2 \rangle = \frac{1}{w} \sum_{i=1}^n w_i x_i^2; \quad w = \sum_{i=1}^n w_i. \quad (3.16)$$

Voor de gewichten geldt dat  $w_i$  evenredig met  $1/\sigma_i^2$  is.

Wanneer een gemiddelde door gewogen middeling van  $x_i \pm \sigma_i$  tot stand is gekomen, dan is de schatting voor de standaarddeviatie in het gemiddelde gelijk aan

$$\hat{\sigma}_{\langle x \rangle} = \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1/2} \quad (3.17)$$

Zie voor uitleg Bijlage G.

Bij deze formule gaan we ervan uit dat de waarden van  $\sigma_i^2$  goed bekend zijn en hebben we de waarde van  $\langle (\Delta x)^2 \rangle$  niet gebruikt voor de schatting van  $\hat{\sigma}_{\langle x \rangle}$ . Of de gevonden spreiding in de meetwaarden statistisch aanvaardbaar is (compatibel met de van tevoren bekende  $\sigma_i^2$ ), is na te gaan met een *chi-kwadraattest* die behandeld wordt in par. 5.4. Voor dit geval is het aantal vrijheidsgraden gelijk aan  $n - 1$  en is

$$\chi^2 = \sum_{i=1}^n \frac{(x_i - \langle x \rangle)^2}{\sigma_i^2} = \frac{\langle (\Delta x)^2 \rangle}{\hat{\sigma}_{\langle x \rangle}^2}. \quad (3.18)$$

De waarde van  $\chi^2$  moet in de buurt van  $n - 1$  liggen. Is  $\sigma_i$  niet goed bekend en zijn er een voldoende aantal meetwaarden, dan kan  $\langle (\Delta x)^2 \rangle$  ook gebruikt worden om  $\hat{\sigma}_{\langle x \rangle}$  te bepalen. Neem dan  $\chi^2 = n - 1$ , zodat

$$\hat{\sigma}_{\langle x \rangle}^2 = \frac{\langle (\Delta x)^2 \rangle}{n - 1}. \quad (3.19)$$

**OPGAVE 3.5** (zie voor antwoorden p. 105)

Stel je zit in de commissie die een nieuwe waarde voor het getal van Avogadro moet bepalen. Je hebt de volgende gegevens, die eerst goed op hun betrouwbaarheid zijn gecontroleerd:

- het al bekende getal (zie het blad FYSISCHE CONSTANTEN, p. 123)
- een meetserie van onderzoeker A die als resultaat geeft:  
 $6,0221348(75) \times 10^{23}$
- een meetserie van onderzoeker B die als resultaat geeft:  
 $6,0221372(30) \times 10^{23}$
- een meetserie van onderzoeker C die als resultaat geeft:  
 $6,02215(5) \times 10^{23}$

Geef het op de juiste wijze gewogen gemiddelde en de standaardfout in het gemiddelde. Voor antwoorden zie blz 105.

**MATHEMATICA VOORBEELD 3.2**

Dit programma **analysew** berekent van een lijst getallen  $x$  met standaarddeviaties  $s$  de beste schattingen voor het gemiddelde en de nauwkeurigheid van het gemiddelde. De waarden van  $\chi^2$  en het aantal vrijheidsgraden worden gegeven en er wordt een chikwadrattest uitgevoerd. (zie sectie 5.4). Zijn gegeven standaarddeviaties onbetrouwbaar of te klein, dan wordt de nauwkeurigheid ook bepaald aan de hand van de afwijkingen van  $x$  zelf. Wordt voor  $s$  een getal i.p.v. een lijst ingevuld, dan worden alle gewichten gelijk aan  $s$  genomen. Er wordt een plotje gemaakt van de punten met hun s.d.'s.

Dit programma gebruikt routines uit een tweetal 'Packages', die eerst moeten worden geladen:

```
<< Statistics'ContinuousDistributions'
<< Graphics'Graphics'
```

```
Off[General::spell1]; ClearAll[analysew];
analysew[x_, s_] := Module[
  {f, n, ns, ss, result, w, wtot, xa, xav, dxsqav, chi2,
   sdav, sdav2},
  n = Length[x]; ns = Length[s];
  If[ns == 0, ss = Table[s, {i, n}], ss = s];
  If[n < 3,
    ((Print[""]; Print["Minder dan 3 punten"]; Return[]));
  ns = Length[ss];
  If[ns != n,
```

```

(Print[""]; Print["x en s ongelijke lengte"];
Return[]]);
w = 1/(ss ss); wtot = Apply[Plus, w]; w = w/wtot;
result = Table[, {4}];
xav = w.x; xa = x - xav; dxsqav = w.(xa xa);
sdav = Sqrt[1/wtot]; chi2 = dxsqav wtot;
sdav2 = Sqrt[dxsqav/(n - 1)];
result[[1]] = xav; result[[2]] = sdav;
result[[3]] = chi2; result[[4]] = sda2;
Print[""]; Print["Resultaat van analysew: Gewogen middeling"];
Print[" gemiddelde = ", N[xav, 6], " +/- ", N[sdav, 6]];
Print[""]; Print["Plot: punten x-xav, met error bars (=s.d.)"];
ErrorListPlot[Transpose[{xa, ss}]];
Print["Test de afwijkingen nu met een chi2 test.
Dit heeft alleen zin als de gegeven waarden van s
betrouwbare standaarddeviaties zijn."];
Print[" chi2 = ", N[chi2, 3], " voor ", N[n - 1, 3],
" vrijheidsgraden"];
f = CDF[ChiSquareDistribution[n - 1], chi2];
Print[" F(chi2) = ", N[f, 3]];
If[f >= 0.9,
Print[" De afwijkingen zijn groter dan verwacht.
Wellicht zijn de sd's in de meetpunten onderschat."],
If[f <= 0.1,
Print[" De afwijkingen zijn kleiner dan verwacht."],
Print[" De afwijkingen zijn redelijk (vallen binnen het
10-90 % betrouwbaarheidsinterval)"]]];
If[f >= 0.5,
Print[" Rekening houdend met de werkelijke afwijkingen
(i.p.v. de gegeven waarden van s) wordt de standaardfout
in het gemiddelde nu ",
N[sdav2, 6]], ];
Return[result];
]

```

### 3.6 Robuuste schattingen

De schattingen van parameters als standaarddeviatie en standaardfout, zoals in secties 3.3 en 3.4 zijn behandeld, zijn betrekkelijk gevoelig voor uitschieters ('outliers') in de meetwaarden. Dat komt doordat de kwadraten van afwijkingen in de verwerking worden gebruikt en een statis-

tische uitschieter dan vrij zwaar meetelt. Is een afwijking groot, dan kan men overwegen zo'n meetwaarde weg te laten (zie hieronder). Ook zijn de gebruikte methoden soms alleen geldig voor normale verdelingen, zoals betrouwbaarheidsintervallen die met de Student's t-methode zijn bepaald. Er zijn in de moderne statistiek z.g. *robuste* methoden ontwikkeld om meetseries te verwerken, waarbij uitschieters een kleinere rol spelen en de resultaten niet van het type verdelingsfunctie afhangen (in het laatste geval spreken we van *distributie-vrije methoden*). Deze methoden berusten op de rangorde van geordende data ('rank-based methods'). We geven hier slechts een summier samenvatting van deze methoden en verwijzen voor verdere details naar de literatuur [5, 13]. In de normale verwerking van meetresultaten is de toepassing van, wat nu heet de *klassieke methoden*, voldoende.

#### *Het weglaten van uitschieters*

Het kan voorkomen dat een meetwaarde veel afwijkt van een verwachte waarde. Dat kan toevallig zijn, maar kan ook het gevolg zijn van een fout of vergissing. Het is verantwoord om zo'n meting weg te laten in de bewerking van de resultaten. Neem hierbij als criterium een afwijking van meer dan  $2,5 \sigma$ . Maar doe dit niet voor meer dan één meting uit een serie. Voorzichtigheid is vereist, omdat het al of niet weglaten van een meting gemakkelijk door subjectieve overwegingen kan worden beïnvloed, als een meetwaarde je niet zo goed uitkomt! Veel beter is het natuurlijk om zo'n meting opnieuw te verrichten: een fout komt dan aan het licht. En wordt er weer een significante afwijking, gevonden, dan is men wellicht op het spoor van een interessant verschijnsel gekomen dat nader onderzoek verdient.

Het criterium van  $2,5 \sigma$  is betrekkelijk willekeurig; sommigen geven de voorkeur aan  $3 \sigma$ . Het criterium moet zo gekozen worden dat het toevallig voorkomen van zo'n afwijking onwaarschijnlijk is, bijv. dat de afwijking minder dan 5% kans heeft om in een meetserie voor te komen. Maar dan is de grens die gesteld moet worden afhankelijk van het *aantal* meetpunten in een serie: de kans dat een enkel meetpunt meer dan  $2,5 \sigma$  afwijkt is ruim 1%; de kans dat minstens één meetpunt in een serie van 20 meetpunten meer dan  $2,5 \sigma$  afwijkt is zelfs meer dan 20%. Het eerste is dus niet erg waarschijnlijk, maar het tweede kan heel goed toevallig voorkomen zonder dat er iets bijzonders met het meetpunt aan de hand is. In de tabel op pag. 2 van het blad NORMALE VERDELING (pag. 133) is de kans weergegeven dat tenminste één meetpunt een bepaalde afwijking overschrijdt in een serie van  $n$  meetpunten. Kiezen we een 5% grens

voor deze kans, dan zien we dat voor minder dan 10 meetpunten  $2,5 \sigma$  een goede keus is; voor 10 tot 50 meetpunten is  $3 \sigma$  beter en voor 50 tot een paar honderd meetpunten is  $3,5 \sigma$  de beste keus.

#### *Rang-gebaseerde schattingen*

Het geschatte gemiddelde van een verdeling wordt normaal berekend als het gemiddelde van de meetwaarden. Men kan daarvoor echter ook de *mediaan* van de verdeling nemen; voor een groot aantal meetpunten komt er hetzelfde uit, maar voor een beperkt aantal meetpunten is de mediaan minder gevoelig voor het voorkomen van uitschieters dan het gemiddelde van de meetwaarden. De mediaan heeft de eigenschap dat er evenveel negatieve als positieve afwijkingen voorkomen en gebruikt dus alleen het *teken* van de afwijkingen.

Een teken-gebaseerde schatting van een betrouwbaarheidsinterval wordt verkregen uit de binomiaalverdeling van het aantal positieve tekens van de mogelijke afwijkingen. Stel we hebben 5 metingen, in opklimmende waarde gesorteerd:  $x_1, x_2, x_3, x_4, x_5$ . We nemen als schatting  $\hat{\mu}$  de mediaan  $x_3$ . Stel dat  $\mu < x_1$ , dan zouden de afwijkingen de tekens +++++ hebben, en de binomiaalkans hierop is

$$p(\mu < x_1) = 2^{-5} \binom{5}{5} = 1/32$$

(zie sectie 2.2). Diezelfde kans krijgen we als  $\mu > x_5$ , dus het interval  $(x_1, x_5)$  heeft een betrouwbaarheidsniveau van  $30/32 = 94\%$ . Ligt  $\mu$  tussen  $x_2$  en  $x_4$ , dan hebben de afwijkingen de tekens ----+ of ---+;

de binomiaalkans hierop is:

$$p(x_2 < \mu < x_4) = 2^{-5} \binom{5}{3} + 2^{-5} \binom{5}{2} = 20/32 = 62\%$$

Omdat we alleen discrete waarden ter beschikking hebben kunnen geen nauwkeurige betrouwbaarheidsniveaus (zoals 90%) worden gegeven. De methode is wel robuust, maar als we een aanwijzing hebben dat we met een normale verdeling te maken hebben, lang niet zo goed als de ‘klassieke’ parameterschattingen.

Tot slot enkele woorden over een andere moderne, ‘distributievrije’ methode, de *bootst rap*, bedoeld om een benaderde waarschijnlijkheidsdistributie (een ‘sampling distributie’) van een gemiddelde te verkrijgen (en daarmee betrouwbaarheidsintervallen) op grond van een meetserie, zonder enige aanname over de verdelingsfunctie waaruit de meetserie is

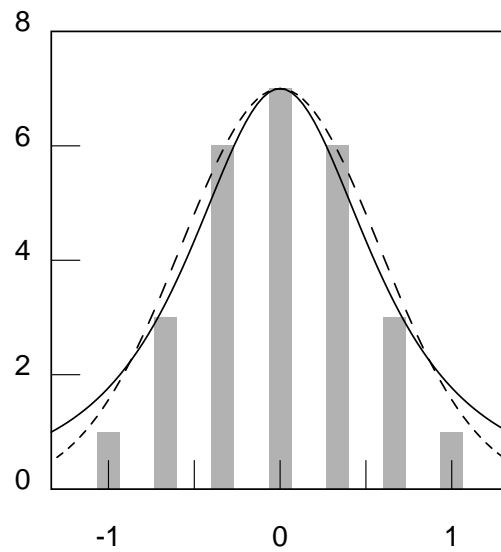
ontstaan.<sup>2</sup> De methode is handig en eenvoudig, maar kan alleen met een computer uitgevoerd worden. Stel we hebben een serie van  $n$  onafhankelijke meetwaarden. Genereer nu een groot aantal (bijv. 10 000) series, ieder van  $n$  ‘metingen’, die elk willekeurig getrokken worden uit de  $n$  oorspronkelijk meetwaarden, maar steeds ‘met teruglegging’, d.w.z. dat na elke trekking de kans op elke meetwaarde hetzelfde blijft. Bepaal van elke serie het gemiddelde. De verzameling van alle zo verkregen gemiddelden benadert de ‘sampling distributie’ van het gemiddelde. Bij een klein aantal metingen kunnen alle mogelijke series worden gegenereerd (dat zijn er  $n^n$ ), maar bij meer dan vijf metingen loopt dat uit de hand. Ter illustratie geven we in figuur 3.4 de bootstrapverdeling weer als we een meetserie van drie waarden: -1, 0 en 1 hebben: er zijn 7 mogelijke gemiddelden. In dezelfde figuur staat ter vergelijking de Student’s  $t$ -distributie voor dezelfde drie meetwaarden (dus twee vrijheidsgraden), waarbij  $\hat{\mu} = 0$  en  $\hat{\sigma} = 1$ . De ‘klassieke’ standaardfout in het gemiddelde is dus  $1/\sqrt{3} = 0.577$ ; de s.d. van de bootstrapverdeling is  $\sqrt{2}/3 = 0,471$ . Ook is de normale verdeling met  $\sigma = 0,577$  weergegeven. We zien dat voor dit symmetrische geval de normale verdeling en de bootstrapverdeling goed overeenkomen; de  $t$ -verdeling is vooral breder aan de voet. Als we maar drie waarden hebben en geen aanwijzing dat de oorspronkelijke verdeling normaal is, is er ook geen goede reden om de  $t$ -verdeling toe te passen.

**OPGAVE 3.6** (zie voor antwoorden p. 105)

Zet de bootstrapverdeling waarvan het histogram in figuur 3.4 is gegeven, uit op waarschijnlijkheidspapier. Voldoet deze aan een normale verdeling? Bepaal grafisch de verwachte waarde en de standaardfout en vergelijk deze met de in de tekst gegeven waarden.

---

<sup>2</sup>*Bootstrap* betekent letterlijk ‘schoenveter’. Er wordt mee bedoeld dat je probeert iets te bereiken dat in principe onmogelijk is, zoals jezelf aan je schoenveters omhoog trekken. Een sampling distributie van een gemiddelde kan je alleen construeren als je veel onafhankelijke meetseries hebt; als je zo’n distributie genereert uit een enkele meetserie houd je jezelf voor de gek. De bootstrapmethode stamt uit 1979, zie B. Efron en R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London, 1993.



Figuur 3.4: Histogram vande bootstrapverdeling van drie meetpunten  $-1$ ,  $0$  en  $1$  (ordinaatwaarden  $\times 1/27$ ). Getrokken lijn: Student's  $t$ -distributie voor 2 vrijheidsgraden; gestreepte lijn: normale verdeling met 'klassieke' standaardfout als s.d. Alle verdelingen zijn geschaald voor gelijke topwaarden.

## Hoofdstuk 4

# De grafische gegevensverwerking en foutendiscussie

### 4.1 Inleiding

In het vorige hoofdstuk hebben we een meetserie bekeken van gelijksoortige meetwaarden  $x_i$ , die bij foutloze meting allemaal hetzelfde hadden moeten opleveren. Vaak worden grootheden echter gemeten als functie van een onafhankelijke variabele, zoals de tijd, de temperatuur of een concentratie. In het algemeen verwacht men op theoretische gronden een zeker verband tussen de gemeten grootheden  $y_i$  en de onafhankelijke variabele  $x_i$  (merk op dat er nu twee reeksen meetwaarden  $x_i$  en  $y_i$  zijn; meestal is  $x_i$  een ingestelde waarde die nauwkeurig bekend is en is  $y_i$  een daarbij horende gemeten waarde, die aan meetfouten onderhevig is). Een voorbeeld van zo'n verband is de lineaire relatie  $y = ax + b$ , maar de relatie kan ook best ingewikkelder zijn, bijv.  $y = c \exp(-kx)$ . Vaak kan een ingewikkelder relatie herleid worden tot een lineair verband (in het laatste geval:  $\ln y = \ln c - kx$ ). Zo'n linearisatie is altijd aan te raden omdat dan grafisch een rechte lijn wordt verkregen en snel kan worden beoordeeld of de meetgegevens aan de verwachte relatie voldoen. In par. 4.2 zullen een aantal voorbeelden worden uitgewerkt.



Keren we terug tot de lineaire relatie  $y = ax + b$ . We meten  $n$  meetpunten  $(x_i, y_i)$ ,  $i = 1 \dots n$ , en we verwachten dat de meetwaarden  $y_i$  zo goed mogelijk voldoen aan de relatie

$$y_i \approx f(x_i) \quad (4.1)$$

waar  $f(x) = ax + b$  de verwachte relatie voorstelt. De opgave die we ons stellen is de parameters  $a$  en  $b$  zo te bepalen dat de meetpunten zo weinig mogelijk van de functiewaarden afwijken.

Maar wat betekent ‘zo goed mogelijk’? De afwijkingen  $\epsilon_i$  tussen meetwaarden en functiewaarden:

$$\epsilon_i = y_i - f(x_i) \quad (4.2)$$

mogen alleen het gevolg zijn van toevallige meetfouten, en we verwachten in het algemeen dat de afwijkingen  $\epsilon_i$  steekproeven zijn uit een normale verdeling met gemiddelde nul. De correcte methode voor dit soort parameteraanpassing is de *kleinste-kwadratenmethode*, die in hoofdstuk 5 wordt beschreven en die normaal met een computer wordt uitgevoerd.

In de praktijk is het lang niet altijd nodig een kleinste kwadrataanpassing uit te voeren of een computerprogramma te gebruiken. Het is altijd zinvol om de meetgegevens zodanig in een grafiek uit te zetten dat een rechte lijn verwacht wordt. Een op het oog zo goed mogelijk getrokken lijn geeft vaak al voldoende nauwkeurige resultaten. Zelfs de fouten in helling en asafsnijding kunnen redelijk goed geschat worden door de lijn binnen de meetfouten te variëren. De computer is nuttig als er veel meetpunten zijn, als de gewichtsfactoren ongelijk zijn of als er grote nauwkeurigheid wordt vereist, maar is geen substituut voor slechte metingen en geeft meestal ook geen dieper inzicht. Integendeel: vaak voor het gebruik van computerprogramma’s die iets doen dat niet goed gedocumenteerd is en dat je niet goed begrijpt.

In dit hoofdstuk bepalen we ons tot eenvoudige grafische verwerking van meetresultaten en de daarbij behorende ‘grove’ foutendiscussie. In veel gevallen is het aan te raden eerst zo’n grove analyse te doen omdat daarmee vaak het beste inzicht wordt verkregen. Daarna kan dan een meer verantwoorde computerverwerking worden toegepast. Een eenvoudig plotje kan het best met de hand worden gemaakt, maar dat kan natuurlijk ook op de computer.

#### MATHEMATICA VOORBEELD 4.1

Een simpel plotje wordt gemaakt met

`ListPlot[{x1, y1}, {x2, y2}, ...]`

Vaak heeft men de lijsten  $x$  en  $y$  apart al bij de hand, en niet in de array-vorm die voor `ListPlot` vereist is. De eenvoudigste manier om  $x$  en  $y$  in de juiste vorm te brengen is de volgende:

`data = Transpose[{x, y}]`

`ListPlot[data]` Dit plot alleen de punten  $x, y$ , maar de plot kan aangevuld worden met functielijnen met het commando `Plot`.

Wil men ook de standaarddeviatie met streepjes aangeven ('errorbars') dan kan dit met het commando `ErrorListPlot`, maar dit commando staat in het pakket `Graphics`Graphics``. Dat moet eerst geladen worden met `<<Graphics`Graphics``.

Daarna kan met

`ErrorListPlot[{{x1, y1, dy1}, {x2, y2, dy2}, ...]`

een plotje gemaakt worden van de punten  $x, y$  met verticale streepjes ter lengte  $\pm dy$ .

## 4.2 Het lineariseren van functies

Om te kunnen beoordelen of een serie meetwaarden aan een verwachte relatie voldoen moeten we proberen de gegevens zo uit te zetten dat de verwachte relatie een rechte lijn oplevert. Zetten we de meetpunten met hun standaardfout in de grafiek en trekken we daar de 'beste' rechte lijn door, dan kan niet alleen worden beoordeeld of de metingen aan de relatie voldoen, maar kunnen de parameters van de lijn ook worden afgelezen uit de grafiek en hun nauwkeurigheden worden afgeschat (par. 4.3).

We geven enkele voorbeelden van linearisaties van functies.

- $y = ae^{-kx}$  :  $\ln y = \ln a - kx$  (voorbeelden: concentratieverloop in de tijd bij eerste-orde reactie, aantal counts per minuut bij radioactief verval). Zet  $\ln y$  op lineaire schaal uit tegen  $x$ , of zet  $y$  op logaritmische schaal uit tegen  $x$ . Gebruik hiervoor enkel-logaritmisch grafiekenpapier (zie pag. ??). Figuur 1.1 op pag. 11 is hiervan een voorbeeld. De helling ( $-k$  in dit voorbeeld) wordt afgelezen door van een niet te klein segment van de lijn de eindpunten  $(x_1, y_1)$  en  $(x_2, y_2)$  af te lezen; de helling is nu  $\ln(y_2/y_1)/(x_2 - x_1)$ . Neem je de punten  $x_1$  en  $x_2$  waarbij de lijn door  $y$ -waarden 1 en 10 gaan (of 10 en 100) dan is de helling eenvoudig  $\ln 10/(x_2 - x_1)$ .
- $y = a + be^{-kx}$  :  $\ln(y - a) = \ln b - kx$ . Schat eerst  $a$  uit de waarden

van  $y$  voor grote  $x$  en zet  $y - a$  op logaritmische schaal uit tegen  $x$ . Als de lijn niet goed recht wordt, corrigeer  $a$  dan een beetje.

- $y = a_1 e^{-k_1 x} + a_2 e^{-k_2 x}$ . Dit gaat niet goed grafisch, tenzij  $k_1$  en  $k_2$  duidelijk verschillen. Ook een computer heeft moeite met dit soort analyses! Bepaal eerst de ‘langzame’ component (kleinste  $k$ ), trek die af van  $y$  en zet het verschil weer logaritmisch uit. Fig. 4.1 geeft het resultaat voor de meetwaarden uit tabel 4.1; de middelbare fout in elke  $y$  is  $\pm 1$  eenheid.

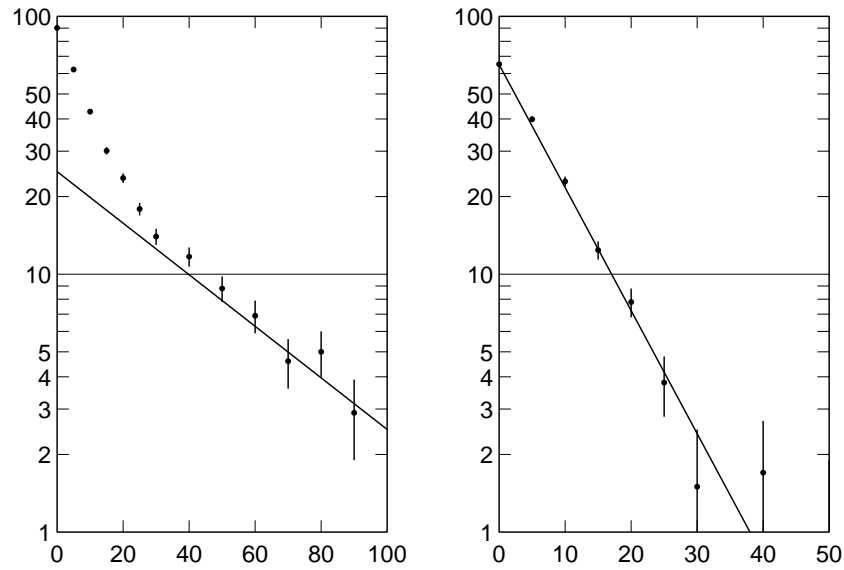
$x$	$y$	$z$	$x$	$y$	$z$
0	90,2	65,2	40	11,7	1,7
5	62,2	39,9	50	8,8	0,9
10	42,7	22,9	60	6,9	0,6
15	30,1	12,4	70	4,6	-0,4
20	23,6	7,8	80	5,0	1,1
25	17,9	3,8	90	2,9	-0,3
30	14,0	1,5			

Tabel 4.1: Meetwaarden  $y$  die een som van twee  $e$ -machten voorstellen. De kolom  $z$  ontstaat na aftrekken van de ‘langzaamste’  $e$ -macht. De middelbare fout in  $y$  is 1 eenheid.

De kolom  $z$  in de tabel geeft de verschillen van  $y$  en de waarden die de getrokken lijn in de linkerfiguur oplevert. Deze lijn is op het oog getrokken en gaat door de punten  $(0, 25)$  en  $(100, 2,5)$ , zodat  $k_2 = [\ln(25/2,5)]/100 = 0,023$ . De vergelijking voor de lijn is dus  $25 \exp(-0.023x)$ . In de rechterfiguur is  $z$  uitgezet; we zien dat de punten redelijk aan een rechte lijn voldoen. De getrokken lijn gaat door de punten  $(0, 65)$  en  $(38, 1)$ , zodat  $k_1 = (\ln 75)/38 = 0,11$ , zodat de functie die alle meetpunten benadert, gelijk is aan

$$f(x) = 65 e^{-0,11 x} + 25 e^{-0,023 x}$$

Het is nogal onbetrouwbaar om op deze grafische wijze een goede foutenschatting te maken in de parameters van deze vergelijking. De grafische bepaling is uitstekend geschikt om te dienen als *beginschatting* van parameters voor een *niet-lineaire kleinste kwadratenaanpassing* die met een computerprogramma moet worden uitgevoerd (zie par. 5.2).



Figuur 4.1: Grafische verwerking van meetgegevens, die een som van twee exponentieel afnemende grootheden zijn. In de linkerfiguur zijn de meetgegevens  $y$  logaritmisch tegen de onafhankelijke variabele  $x$  uitgezet en wordt de 'langzaamste' component met een rechte lijn benaderd. In de rechterfiguur zijn de verschillen  $z$  tussen de meetwaarden en de 'langzame' component uitgezet. Let op de verschillende schaal van  $x$ .

- $y = (x - a)^p$  (voorbeeld: de isotherme compressibiliteit  $\chi$  van een vloeistof in de buurt van de kritische temperatuur  $T_c$  gedraagt zich als functie van de temperatuur als  $\chi = C(T - T_c)^{-\gamma}$ , waar  $\gamma$  de *kritische exponent* is). We zetten  $\log y$  tegen  $\log(x - a)$  uit (dus  $y$  tegen  $(x - a)$  op een dubbellogaritmische schaal); als  $a$  niet bekend is, variëren we  $a$  enigszins tot de lijn recht is. De helling van de lijn geeft  $p$ .
- $y = ax/(b + x)$  (voorbeelden: geadsorbeerde hoeveelheid stof  $n_{ads}$  tegen concentratie  $c$  in oplossing of druk  $p$  in de gasfase bij Langmuir adsorptie:  $n_{ads} = n_{max}c/(K + c)$ ; reactiesnelheid  $v$  als functie van substraatconcentratie  $[S]$  bij Michaelis-Menten enzymkinetiek:  $v = v_{max}[S]/(K_m + [S])$ ). Deze vergelijking draaien we om en krijgen nu een lineaire relatie tussen  $1/y$  en  $1/x$ :

$$\frac{1}{y} = \frac{1}{a} + \frac{b}{a} \frac{1}{x} \quad (4.3)$$

In de enzymkinetiek heet een grafiek van  $1/v$  tegen  $1/[S]$  een *Lineweaver-Burk plot*<sup>1</sup>. Er zijn nog twee andere methoden die ook een rechte lijn opleveren: zet  $x/y$  uit tegen  $x$  (de *methode van Hanes*):

$$\frac{x}{y} = \frac{b}{a} + \frac{x}{a} \quad (4.4)$$

of zet  $y/x$  uit tegen  $y$  (de *methode van Eadie-Hofstee*):

$$\frac{y}{x} = \frac{a}{b} - \frac{y}{b} \quad (4.5)$$

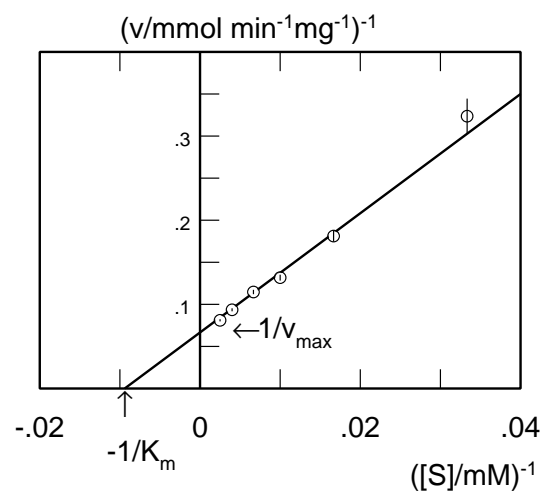
Met de experimentele waarden voor de snelheid van de omzetting  $v = y$  van ureum door het enzym urease<sup>2</sup> als functie van de ureumconcentratie  $[S] = x$  uit tabel 4.2 worden de grafieken van fig. 4.2 en fig. 4.3 verkregen. In een Lineweaver-Burk plot kan de waarde van  $K_m = b$  uit de horizontale asafsnede van de geschatte lijn worden afgelezen en de waarde van  $v_{max} = a$  uit de verticale asafsnede. Het afschatten van nauwkeurigheden van de parameters uit deze grafieken is niet erg betrouwbaar: ook hier is het beter een niet-lineaire kleinste kwadratenaanpassing met de computer uit te voeren, waarbij de grafische schattingen goede beginwaarden zijn.

<sup>1</sup>zie bijv. N.C. Price and R.A. Dwek, *Principles and problems in physical chemistry for biochemists*, Clarendon, Oxford, 1979.

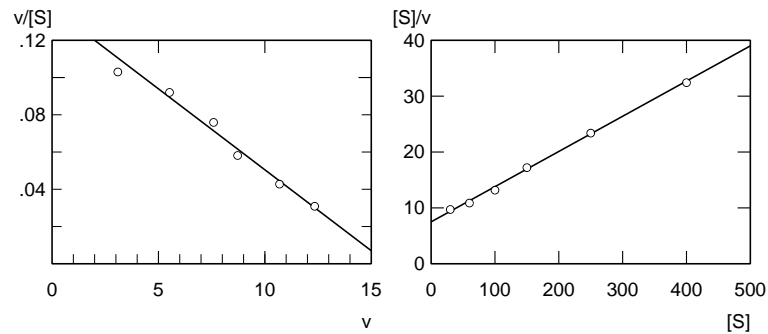
<sup>2</sup>Voorbeeld uit Price en Dwek, met additionele ruis.

[S] mM	1/[S] 1/mM	$v$ mmol min <sup>-1</sup> mg <sup>-1</sup>	$\sigma_v$ mmol min <sup>-1</sup> mg <sup>-1</sup>	1/ $v$ mmol <sup>-1</sup> min mg	$\sigma_{1/v}$ mmol <sup>-1</sup> min mg
30	0,03333	3,09	0,2	0,3236	0,0209
60	0,01667	5,52	0,2	0,1812	0,0066
100	0,01000	7,59	0,2	0,1318	0,0035
150	0,00667	8,72	0,2	0,1147	0,0026
250	0,00400	10,69	0,2	0,09355	0,0018
400	0,00250	12,34	0,2	0,08104	0,0013

Tabel 4.2: Snelheid  $v$  van de omzetting van ureum door het enzym urease, als functie van de ureumconcentratie  $[S]$ . Ten behoeve van een Lineweaver-Burk plot zijn ook de inverse waarden gegeven. De s.d. van  $1/v$  is gelijk aan  $\sigma_v/v^2$ .



Figuur 4.2: Een Lineweaver-Burk plot van de tabelgegevens.



Figuur 4.3: Eadie-Hofstee (links) en Hanes (rechts) plot van de tabelgegevens.

**OPGAVE 4.1** Trek een rechte lijn door de punten van figuur 1.1 en bepaal de parameters in  $c(t) = c_0 e^{-kt}$ .

**OPGAVE 4.2** Bepaal uit figuren 4.2 en 4.3 de waarden van  $v_{max}$  en  $K_m$ . De op het oog getrokken lijnen gaan door de punten  $(-0.0094, 0)$  en  $(0,04, 0,35)$  (Lineweaver-Burk),  $(0,04, 0,35)$  en  $(15, 0,007)$  (Eadie-Hofstee) en  $(0, 7,5)$  en  $(500, 39)$  (Hanes).

### 4.3 Grafische schattingen van de nauwkeurigheid van parameters in lineaire functies

In de vorige paragraaf werd aangeven hoe meetgegevens grafisch kunnen worden uitgezet en hoe de parameters van een functie geschat kunnen worden uit de beste lijn die door de punten kan worden getrokken. We zullen nu aangeven hoe uit de grafiek ook een eenvoudige schatting van de nauwkeurigheid van de parameters kan worden gemaakt. Soms zijn zulke schattingen voor de praktijk voldoende. In het volgende hoofdstuk komt dan een nauwkeuriger statistische procedure (de kleinste kwadratenmethode) aan de orde, die meestal met een computerprogramma wordt uitgevoerd.

Voor een foutenschatting is het noodzakelijk ook de meetfouten in de grafiek op te nemen. Is de fout in de onafhankelijk variabele  $x$  verwaarloosbaar, dan volstaat het geven van een vertikaal foutenstreepje van

$y - \sigma_y$  tot  $y + \sigma_y$ . Is de fout in  $x$  ook van belang, dan moet ook een horizontaal foutenstreepje van  $x - \sigma_x$  tot  $x + \sigma_x$  worden opgenomen. Een duidelijke weergave is een *ellips* met als hoofdassen  $2\sigma_x$  en  $2\sigma_y$ .

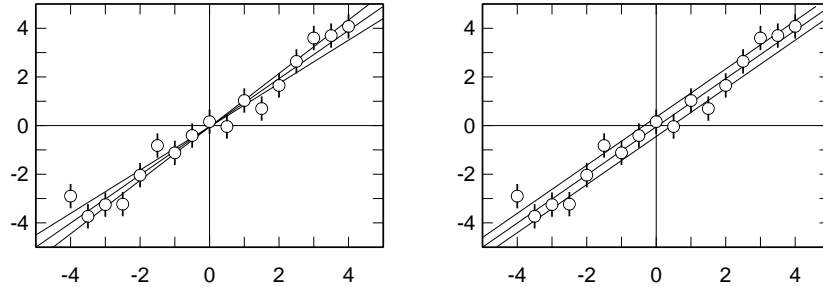
De beste rechte lijn door de meetpunten sluit zo goed mogelijk aan bij alle punten  $(x_i, y_i)$ . Als maatstaf kan worden genomen dat de som van de afstanden (inclusief teken) van de punten tot de rechte gelijk nul is. Maar dit bepaalt de lijn nog niet: elke lijn door het *zwaartepunt* van de meetpunten ( $\langle x \rangle, \langle y \rangle$ ) voldoet hieraan, zodat de helling hiermee nog niet is vastgelegd. Nu kan men op het oog de lijn trekken die zo goed mogelijk door alle punten gaat. Men kan deze lijn ook construeren met de *zwaartepuntsmethode*: de waarnemingsparen worden in twee even grote groepen verdeeld, een linker en een rechter, en de gevraagde rechte wordt door de zwaartepunten van beide groepen getrokken. Uit de zo getrokken beste lijn kunnen de twee parameters  $a$  en  $b$  bepaald worden uit de helling en de asafsnede. Het kan zijn dat de asafsnede  $b$  voor  $x = 0$  niet binnen de grafiek valt; in dat geval moet  $b$  worden geëxtrapoleerd uit de afsnede van de verticale as van de grafiek, met gebruikmaking van de waarde van  $a$  die uit de helling is bepaald.

Nadat de beste lijn bepaald is, kan eerst worden gekeken of de meetwaarden aan de verwachte (lineaire) functie voldoen. Als de foutenstreepjes of ellipsjes inderdaad de middelbare of standaardfout ('rms error') voorstellen dan moet de lijn door ongeveer 2/3 van de meetpunten inclusief foutengrenzen gaan (zie tabel 2.1 op pag. 24). Dus een derde van de punten mag er (een beetje) buiten vallen. Er kunnen ook uitschieters tussen de data vallen, die extra aandacht vereisen (zie p. 41: weglaten of opnieuw meten?)

Vervolgens kan een ruwe schatting van de fouten in de parameters worden verkregen door de lijn binnen de meetfouten te variëren. Daarbij moet de lijn ongeveer 2/3 van de meetpunten binnen de fout blijven omvatten. De helling  $a$  en de asafsnede  $b$  zijn in het algemeen van elkaar afhankelijk; alleen als het nulpunt van  $x$  in het zwaartepunt wordt gekozen, zijn  $a$  en  $b$  onafhankelijk van elkaar. Daarom moet bij het afschatten van de onnauwkeurigheid van de *helling* de lijn (ongeveer) om het zwaartepunt van de meetpunten worden gedraaid; het beste kan voor de vorm van de lineaire relatie  $y = a(x - \langle x \rangle) + b$  worden gekozen omdat  $a$  en  $b$  dan onafhankelijk zijn. Op en neer schuiven van de lijn geeft een schatting voor de onnauwkeurigheid van  $b$ . Figuur 4.4 geeft een voorbeeld. De helling in de linkerfiguur varieert 10 % ( $a = 1,0 \pm 0,1$ ) en de constante varieert 0,4 ( $b = 0,0 \pm 0,4$ ). Het aantal meetpunten dat buiten de lijnen



valt is 6 tot 8 van de 17. Deze foutenschatting is zeer onnauwkeurig en een betere computeranalyse is aan te bevelen. Men heeft de neiging op deze grafische manier fouten te overschatten.



Figuur 4.4: Het schatten van de fout in de helling  $a$  van een lijn  $ax + b$  (links) en de additieve constante  $b$  (rechts) door zodanige variatie dat de lijn ruwweg  $2/3$  van de meetpunten blijft omvatten. Het zwaartepunt  $(\langle x \rangle, \langle y \rangle)$  is als nulpunt gekozen.

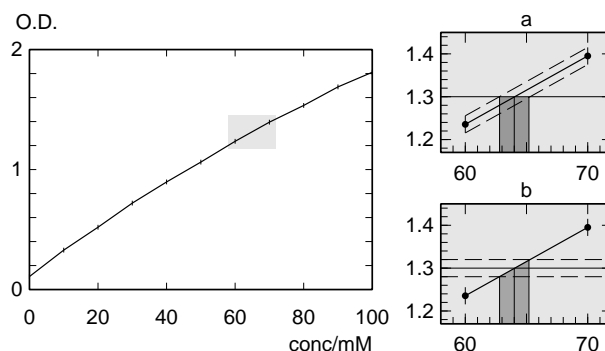
**OPGAVE 4.3** Trek de beste rechte door de meetpunten van de grafiek die in opgave 1.3 is gemaakt. Bepaal de constanten  $A$  en  $E$  in de relatie  $k = A \exp(-E/RT)$  (denk aan de eenheden). Maak een grafische schatting van de fout in  $A$  en  $E$ .

## 4.4 Meetgegevens als ijklijn

Stel dat we met een instrument of meetmethode werken, waarbij uit een meetresultaat  $y$  (bijv. de aflezing van een meter) een grootte  $x$  (bijv. een concentratie) moet worden afgeleid. Wanneer het instrument niet (of niet betrouwbaar) *geijkt* is (d.w.z. dat de uitlezing van het instrument niet direct de te bepalen grootte weergeeft), dan maken we eerst een *ijkkromme*, door bij een aantal goed bekende concentraties  $x_i$  de meetresultaten  $y_i$  te bepalen en deze in een grafiek uit te zetten. De onbekende concentratie  $x$  van een bepaling  $y$  kan dan worden teruggezocht in de ijkcurve. Hoe gaan we hierbij te werk en hoe bepalen we de fout in  $x$ ?

Zorg er bij de ijking voor dat er ijkpunten zijn over het hele gebied dat voor metingen zal worden gebruikt. Extrapolatie buiten het ijkgebied is onbetrouwbaar, maar het is ook niet nodig het instrument te ijken voor

meetwaarden die nooit aan de orde zullen komen. Bepaal weer de beste rechte door de meetpunten die als ijking zijn gebruikt (de ijkljn kan natuurlijk ook een kromme lijn zijn; dan geldt de beschouwing van de fout voor een klein gedeelte van de ijkkromme dat bij benadering recht mag worden genomen). Trek nu twee lijnen parallel aan de ijkljn, zo dat die door de meeste (zeg: tweederde) van de ellipsjes van de meetpunten gaan. Zie figuur 4.5.



Figuur 4.5: Voorbeeld van een ijkljn voor spectrometrische concentratiebepaling: de optische dichtheid  $O.D. = \log(\text{invallende intensiteit}/\text{doorgelaten intensiteit})$  als functie van de concentratie van een stof. Het grijs aangegeven gebiedje is rechts uitvergroet weergegeven: (a) de ijkfout in de concentratie, (b) de fout in de concentratie als gevolg van de onnauwkeurigheid in de gemeten  $O.D.$

Wordt nu een waarde  $y$  gemeten, dan kan de bijbehorende  $x$ -waarde in de grafiek worden teruggelezen van het snijpunt van de horizontale lijn ter hoogte  $y$  en de ijkljn. De foutengrenzen in  $x$  ten gevolge van de onnauwkeurigheid van de ijkljn kunnen nu worden teruggelezen van de snijpunten met de twee lijnen die parallel aan de ijkljnen zijn getrokken (zie fig. 4.5 a). De onnauwkeurigheid van  $y$  zelf leidt ook tot een fout in  $x$ , die rechtstreeks van de ijkljn zelf kan worden afgelezen (fig. 4.5 b). Beide fouten tellen in principe kwadratisch op. Als de ijking zorgvuldig is uitgevoerd, dan zal de ijkfout meestal kleiner zijn dan de fout in  $x$  ten gevolge van de onnauwkeurigheid in de meetwaarde  $y$ . Eén ijkljn volstaat dan en er kan met een schatting van de fout ten gevolge van de onnauwkeurigheid in de meetwaarde worden volstaan.

De hier vermelde grafische methoden geven slechts ruwe schattingen van de fouten, maar deze zijn in de praktijk vaak voldoende, vooral als het

om een klein aantal meetpunten gaat. Statistisch nauwkeuriger is het om een *ijkfunctie* door een kleinste kwadratenaanpassing uit de ijkgegevens af te leiden (zie hoofdstuk 5) en deze te gebruiken om een meetgegeven (inclusief nauwkeurigheid) om te rekenen tot de gevraagde grootte. In plaats van een ijkfunctie kan ook een *ijktabel* of een *correctietabel* worden gebruikt, die voor gebruik geïnterpoleerd moet worden.

**OPGAVE 4.4** Bepaal aan de hand van fig. 4.5 de concentratie met s.d. als de gemeten optische dichtheid gelijk is aan  $1,38 \pm 0,01$ , onder de aanname dat de ijkfout zelf verwaarloosbaar is.

## Hoofdstuk 5

# Kleinste- kwadratenaanpassing en lineaire regressie

### 5.1 Lineaire regressie

Wanneer een stel meetwaarden  $(x_i, y_i)$ ,  $i = 1, \dots, n$  zo goed mogelijk aan een functie  $y = f(x)$  moet voldoen, en de afwijkingen zijn onafhankelijk en voldoen aan een normale verdeling, dan wordt de beste aanpassing verkregen wanneer *de gewogen som van de kwadratische afwijkingen minimaal is*. Of: de parameters in de functie  $f(x)$  moeten zo bepaald worden dat

$$S = \sum_{i=1}^n w_i (y_i - f_i)^2 \text{ minimaal,} \quad (5.1)$$

waarbij  $f_i = f(x_i)$ . Hierbij nemen we aan dat de fout in de waarden van  $x$  verwaarloosbaar is. Hoe te handelen als niet alleen  $y$ , maar ook  $x$  een niet te verwaarlozen fout heeft, komt hieronder aan de orde.

De factoren  $w_i$  zijn de *gewichten* van de meetpunten. Het komt vaak voor dat alle waarnemingen statistisch vergelijkbaar zijn en de gewichten alle  $= 1$  genomen kunnen worden. Zijn de gewichten  $w_i$  ongelijk (doordat de meetpunten zelf een verschillende standaarddeviatie  $\sigma_i$  hebben), dan moeten de gewichten  $w_i$  gelijk aan of evenredig met  $1/\sigma_i^2$  (dus niet met

$1/\sigma$ ) genomen worden. De met  $1/\sigma_i^2$  gewogen som van kwadratische afwijkingen wordt aangeduid met  $\chi^2$  (chi kwadraat):

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - f_i)^2}{\sigma_i^2} \quad (5.2)$$

Zijn de fouten in  $x$  niet verwaarloosbaar (maar wel onafhankelijk van de fouten in  $y$ ) dan moet de  $\sigma_i^2$  in de formule voor  $\chi^2$  vervangen worden door:

$$\sigma_i^2 = \sigma_{y_i}^2 + \left( \frac{\partial f}{\partial x} \right)_{x=x_i}^2 \sigma_{x_i}^2 \quad (5.3)$$

omdat het eigenlijk om de fout in de grootte  $y_i - f(x_i)$  gaat.

Het bepalen van deze *kleinste kwadratenoplossing* voor de parameters in de functie gebeurt meestal met een computerprogramma. Alleen voor *lineaire regressie*, dat is aanpassing aan een lineaire functie van de parameters, kunnen er expliciete formules gegeven worden. Zonder bewijs (zie voor verdere details Bijlage H) volgen hier de formules voor de veel voorkomende functie

$$f = ax + b \quad (5.4)$$

Deze functie is slechts een voorbeeld van lineaire regressie; niet het feit dat de functie lineair is in  $x$  (zoals  $ax + b$ ) is belangrijk, maar elke functie die lineair is in de *parameters* (dus ook  $ax^2 + bx + c$  of  $a + b \log x + c/x$ ) kan door lineaire regressie worden aangepast. Voor het algemene geval moet een stelsel van lineaire vergelijkingen (met de computer) worden opgelost. Zie voor de algemene formules het blad KLEINSTE KWADRATENAANPASSING.

Voor  $f = ax + b$  en niet-verwaarloosbare fouten in  $x$  geldt

$$\sigma_i^2 = \sigma_{y_i}^2 + a^2 \sigma_{x_i}^2 \quad (5.5)$$

Om deze  $\sigma$ 's te berekenen, moet de helling  $a$  bekend zijn. Maar deze moeten we juist nog bepalen! Een probleem is dit niet, want een grove schatting van  $a$  is voldoende en die kan gemakkelijk uit een grafiekje worden bepaald.

De parameters  $a$  en  $b$  volgen uit een aantal gemiddelden van de meetwaarden. Zijn de gewichtsfactoren ongelijk, dan moeten de gewichten bij het middelen worden meegenomen, zoals in par 3.5 is vermeld. Zo is bijvoorbeeld

$$\langle xy \rangle = \frac{1}{w} \sum_{i=1}^n w_i x_i y_i; \quad w = \sum_{i=1}^n w_i, \quad (5.6)$$

waarbij de weegfactoren  $w_i$  evenredig met  $1/\sigma_i^2$  moeten zijn.

$$a = \frac{\langle(\Delta x)(\Delta y)\rangle}{\langle(\Delta x)^2\rangle}; \quad b = \langle y \rangle - a\langle x \rangle, \quad (5.7)$$

waar

$$\Delta x = x - \langle x \rangle; \quad \Delta y = y - \langle y \rangle. \quad (5.8)$$

De gemiddelden van  $x$  en  $y$  hoeven niet eerst afgetrokken te worden, omdat

$$\langle(\Delta x)(\Delta y)\rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle; \quad (5.9)$$

$$\langle(\Delta x)^2\rangle = \langle x^2 \rangle - \langle x \rangle^2. \quad (5.10)$$

Let evenwel bij grote getallen op de numerieke precisie (zie opmerking op pag. 32).

Uit de formule voor  $b$  zien we dat de functie door het punt  $(\langle x \rangle, \langle y \rangle)$  gaat. Dat is het *zwaartepunt* van de verzameling meetpunten. Dit feit hebben we al gebruikt bij de grafische foutendiscussie in hoofdstuk 4.

De schattingen voor de standaarddeviaties  $\sigma_a$  en  $\sigma_b$  van  $a$  en  $b$  kunnen bepaald worden als de wortel uit de geschatte *varianties*  $\hat{\sigma}_a^2$  en  $\hat{\sigma}_b^2$ :

$$\hat{\sigma}_a^2 = \frac{S}{w(n-2)\langle(\Delta x)^2\rangle}; \quad (5.11)$$

$$\hat{\sigma}_b^2 = \hat{\sigma}_a^2 \langle x^2 \rangle, \quad (5.12)$$

waar  $w$  het totale gewicht van alle meetpunten samen is. Zijn alle gewichten gelijk aan 1 genomen, dan is  $w$  dus eenvoudig gelijk aan  $n$ .

Bovenstaande formules voor de varianties van  $a$  en  $b$  gelden alleen als de meetwaarden statistisch onafhankelijk van elkaar zijn. Zijn ze afhankelijk (het kan bijvoorbeeld heel goed voorkomen dat bij tijdseries de opvolgende metingen met elkaar gecorreleerd zijn door een lange tijdsconstante van het meetapparaat) dan is de kleinste kwadratenprocedure nog wel geldig om de beste parameters te bepalen, maar de varianties zijn groter dan hierboven aangegeven. Het aantal meetpunten moet dan gedeeld worden door de *correlatielengte* van de meetserie. Zie de discussie in par. 3.4 en Bijlage F.

De  $n - 2$  in formule 5.11 heeft de betekenis van het aantal ‘vrijheidsgraden:’ het aantal meetpunten min het aantal parameters in de functie. We zien dat  $n$  tenminste drie moet zijn om een variantie te kunnen schatten. Dat is ook logisch, want door twee punten kun je altijd een rechte

lijn trekken;  $S$  is dan nul en er zijn geen gegevens over om een variantie te schatten.

Hebben de meetpunten *van te voren goed bekende standaarddeviaties*, dan bestaat er een manier om na te gaan of de aanpassing aan de meetpunten acceptabel is. Dit heet de *chi-kwadraattest*, die in par. 5.4 wordt behandeld. Zo'n test kan zinvol zijn wanneer men een keus moet maken uit verschillende theoretische functies.

Naast de varianties in  $a$  en  $b$ , die de standaarddeviaties (de middelbare fouten) in  $a$  en  $b$  bepalen, is er ook een grootheid die de *covariantie* van  $a$  en  $b$  wordt genoemd. Deze geeft aan in hoeverre  $a$  en  $b$  met elkaar gecorreleerd zijn, of: in hoeverre een afwijking in  $a$  door een afwijking in  $b$  gecompenseerd kan worden. Covarianties moeten gebruikt worden om de fout te bepalen in grootheden die een functie zijn van de aangepaste parameters, bijvoorbeeld als meetwaarden geïnterpoleerd of geëxtrapoleerd worden (zie Bijlage H). De covariantie is gegeven door

$$\text{cov}(a, b) = -\hat{\sigma}_a^2 \langle x \rangle \quad (5.13)$$

De covariantie wordt ook vaak gegeven relatief t.o.v. het product  $\hat{\sigma}_a \hat{\sigma}_b$  en heet dan *correlatiecoëfficiënt tussen  $a$  en  $b$*  (een dimensieloos getal tussen -1 en +1):

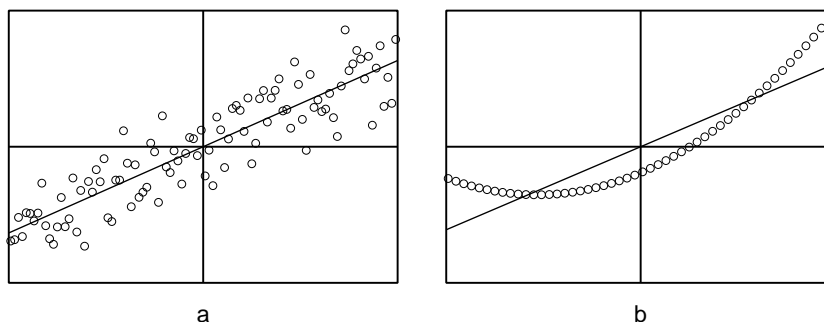
$$\rho_{ab} = \frac{\text{cov}(a, b)}{\hat{\sigma}_a \hat{\sigma}_b} = -\frac{\langle x \rangle}{\sqrt{\langle x^2 \rangle}} \quad (5.14)$$

Merk op dat  $a$  en  $b$  ongecorreleerd zijn wanneer  $\langle x \rangle = 0$ , dus wanneer het nulpunt van  $x$  in het zwaartepunt van de  $x$ -waarden wordt gekozen. Ook dit feit hebben we al benut bij de grafische foutendiscussie. Zo'n keuze is handig wanneer men op grond van een kleinste kwadratenaanpassing een extrapolatie wil maken, met een schatting van de onnauwkeurigheid. Bij de formule  $f(x) = ax + b$  kan de fout dan eenvoudig als de wortel uit de som van kwadratische fouten van  $ax$  en  $b$  worden berekend. Is de relatie tussen  $a$  en  $b$  niet nul, dan moet een correctie worden aangebracht (zie Bijlage A):

$$\sigma_f^2 = \sigma_a^2 x^2 + \sigma_b^2 + 2\rho_{ab}\sigma_a\sigma_b x. \quad (5.15)$$

Er bestaat een manier om na te gaan hoe goed de meetpunten op een rechte liggen. Hiertoe berekent men de *correlatiecoëfficiënt van de serie punten  $r$* , die ook wel de *regressiecoëfficiënt* wordt genoemd (niet te verwarren met de hierboven behandelde correlatiecoëfficiënt van  $a$  en  $b$ ):

$$r = \frac{\langle (\Delta x)(\Delta y) \rangle}{\sqrt{\langle (\Delta x)^2 \rangle \langle (\Delta y)^2 \rangle}} \quad (5.16)$$



Figuur 5.1: Twee puntenverzamelingen die dezelfde regressiecoëfficiënt (correlatiecoëfficiënt tussen  $x$  en  $y$ -waarden van de meetpunten)  $r = 0,900$  hebben. Er zijn geen getallen langs de assen gezet: correlatiecoëfficiënten zijn onafhankelijk van lineaire schaling of translatie van de assen. De getrokken lijnen zijn de beste lineaire fit door de punten.

$$= a \sqrt{\frac{\langle(\Delta x)^2\rangle}{\langle(\Delta y)^2\rangle}} \quad (5.17)$$

Als  $r = \pm 1$ , dan is er een volledige correlatie: de punten liggen precies op een rechte lijn; als  $r = 0$  dan is er geen enkele correlatie en heeft aanpassing aan een rechte lijn geen zin. Voor een redelijk goede correlatie moet  $r$  boven 0,9 liggen. Soms wordt  $r^2$  in plaats van  $r$  gegeven.

De regressiecoëfficiënt geeft aan hoe goed de meetpunten op een rechte lijn liggen, maar geeft niet aan van welk type de afwijkingen zijn. Zo hebben de verzamelingen punten van fig. 5.1a en fig. 5.1b dezelfde regressiecoëfficiënt van 0,900; beide wijken op heel verschillende wijze van de best gefitte rechte lijn af. Uit deze figuur is ook te zien dat  $r = 0,9$  helemaal nog geen goede overeenkomst met een rechte lijn hoeft te betekenen.

Veel wetenschappelijke zakrekenmachines hebben een ingebouwd programma voor lineaire regressie van de functie  $ax + b$ . Dit maakt meestal geen gebruik van gewichtsfactoren; wees hierop verdacht! Let er ook op of  $r$  of  $r^2$  als regressiecoëfficiënt wordt gegeven.

#### MATHEMATICA VOORBEELD 5.1

Het volgende programma `linreg` voert een gewogen lineaire regressie uit op  $y \pm s \approx ax + b$ . Invoer zijn drie lijsten  $x, y, s$ ; de uitvoer wordt afgedrukt, maar is ook beschikbaar als de lijst



```
linreg = {a, b, r, siga, sigb, rhoab}.
```

Het programma geeft `chisq`: de waarde van de gewogen som van kwadratische afwijkingen, en `nu`: het aantal *vrijheidsgraden*. Deze worden dan gebruikt in een *chi-kwadraattest* (zie par. 5.4) om vast te stellen of de aanpassing redelijk is. De s.d. in de parameters *a* en *b* worden op twee manieren uitgerekend:

1. `siga` en `sigb`: volgens vgl 5.11 en 5.12; deze methode gebruikt de spreiding van de meetresultaten en gaat er van uit dat de opgegeven standaarddeviaties in de meetwaarden niet betrouwbaar zijn (in dat geval is ook een chi-kwadraattest niet zinvol).

2. Met gebruikmaking van de opgegeven standaarddeviaties in de meetwaarden. Dit is zinvol als de chi-kwadraattest een aanvaardbaar resultaat geeft en de opgegeven s.d.'s betrouwbaar zijn. Zie de discussie in par. 5.4.

Ook voor dit programma moeten twee 'packages' geladen zijn:

```
<< Statistics'ContinuousDistributions'
<< Graphics'Graphics'
```

```
Off[General::spell1]; ClearAll[linreg];
linreg[x_,y_,s_] := Module[
  {a,b,chisq,data,dx,dxdyav,dxsqav,dy,dysqav,f,i,n,ns,nu,siga,
   sigb,sig2a,sig2b,rhoab,r,w,wtot,xav,xsqav,yav,yminf,ysqav},
  n=Length[x]; ns=Length[s];
  If[ns == 0, ss=Table[s,{i,n}], ss=s];
  If[(Length[y] != n || Length[ss] != n),
    (Print[""];
     Print["Lijsten van ongelijke lengte in linreg"];
     Return[])];
  If[n<3,
    (Print[""]; Print["Niet genoeg punten in linreg"]);
    Return[])];
  w=1/(ss ss); wtot=Apply[Plus,w]; w=w/wtot;
  xav=w.x; yav=w.y;
  dx=x-xav; dy=y-yav;
  xsqav=w.(x x); ysqav=w.(y y);
  dxsqav=w.(dx dx); dysqav=w.(dy dy);
  dxdyav=w.(dx dy);
  a=dxdyav/dxsqav; b=yav-a xav;
  yminf=y-a x-b;
  chisq=w.(yminf yminf) wtot; nu=n-2;
  siga=Sqrt[chisq/(wtot nu dxsqav)];
  sigb=siga Sqrt[xsqav];
  sig2a=1./Sqrt[wtot dxsqav];
```

```

sig2b=sig2a Sqrt[xsqav];
rhoab=-xav/Sqrt[xsqav];
r=dx dyav/Sqrt[dxsqav dysqav];
f=CDF[ChiSquareDistribution[nu],chisq];
Print[""];
Print["Resultaat van linreg: Aanpassing van y op ax+b."];
Print["regressiecoefficient (= correlatiecoëff tussen x en y)"];
Print["= ",N[r,3]];
Print["chisquare = ",chisq," voor ",nu," vrijheidsgraden"];
Print["F(chisq) = ", N[f,3]," (0.1 - 0.9 is acceptabel)"];
Print[""];
Print["a = ",N[a,4],"; b = ",N[b,4]];
Print[""];
Print["als F(chisq)<0.5, gebruik dan voor de s.d. van a en b"];
Print["de waarden gebaseerd op de gegeven s.d. in y:"];
Print["s.d. (a) = ",N[sig2a,3],"; s.d. (b) = ",N[sig2b,3]];
Print[""];
Print["als F(chisq)>0.5, gebruik dan voor de s.d. van a en b"];
Print["de waarden gebaseerd op de spreiding in y:"];
Print["s.d. (a) = ",N[siga,3],"; s.d. (b) = ",N[sigb,3]];
Print[""];
Print["correlatiecoëff tussen a en b = ",N[rhoab,3]];
Print[""];
data = Table[{x[[i]],y[[i]],ss[[i]]},{i,n}];
DisplayTogether[ErrorListPlot[data],
  Plot[a t + b,{t,Min[x],Max[x]}]];
Print[""];
Print[" Plot de residuen:"];
data = Table[{x[[i]],yminf[[i]],ss[[i]]},{i,n}];
ErrorListPlot[data];
Return[{a,b,r,siga,sigb,rhoab}];
]

```

**OPGAVE 5.1** Pas `linreg` toe op de enzymkinetiekgegevens van Tabel 4.2 (pag. 51). Doe dit volgens de Lineweaver-Burk plot, dus  $x = 1/[S]$  en  $y = 1/v$ . Gebruik de goede sigma's voor  $y$ . Geef waarden voor  $v_{max}$  en  $K_m$  met hun standaarddeviaties (denk er aan dat voor de s.d. in een combinatie van  $a$  en  $b$  ook de correlatiecoëfficiënt tussen  $a$  en  $b$  nodig is). Vergelijk die waarden met de grafische schattingen van figuur 4.2. Maak met MATHEMATICA een plotje van de meetpunten en de gefitte lijn.

## MATHEMATICA VOORBEELD 5.2

Wanneer alle meetpunten gelijk gewicht hebben, kan voor lineaire

regressie ook de opdracht `Fit` gebruikt worden: `Fit[data, functies, variabele(n)]` bepaalt de kleinste kwadratenaanpassing van een lineaire combinatie van *functies* aan *data*. Deze procedure is niet beperkt tot  $f(x) = ax + b$ , maar kan elke lineaire functie van de parameters aan. `Fit` is bedoeld voor een snelle bepaling, zonder weegfactoren, en zonder dat de varianties en covarianties van de parameters worden uitgerekend. Voorbeelden:

Pas  $ax + b$  aan op een lijst gegevens  $y = \{y_1, y_2, \dots, y_n\}$ :

```
Fit[y, {x, 1}, x]
```

geeft als uitvoer  $ax + b$ , waarbij  $x$  loopt van 1 tot  $n$ . Voor *functies* staat hier ingevuld  $\{x, 1\}$  voor de vorm  $ax + b$ ; voor aanpassing aan  $ax^2 + bx + c$  zou er  $\{x^2, x, 1\}$  moeten staan.

Pas  $ax + b$  aan op gegevens  $y_1, y_2, \dots, y_n$  die met de functie  $ax + b$  op de  $x$ -waarden  $x_1, x_2, \dots, x_n$  overeen moet komen:

```
Fit[{{x1, y1}, ..., {xn, yn}}, {x, 1}, x]
```

Als de getallen  $x$  in een list *xlist* en  $y$  in een andere list *ylist* staan, dan gaat het zo:

```
Fit[Transpose[{xlist, ylist}], {x, 1}, x]
```

Denk er aan dat de variabele naam (hier  $x$ ) niet gelijk mag zijn aan de naam van een van de lists (zoals *xlist*). Heten de lijsten  $x$  en  $y$  en is een plot van de data en de gefitte lijn voor  $xmin < x < xmax$  nodig, dan doen de volgende regels de job:

```
data = Transpose[{x,y}]
fun = Fit[data, {x, 1}, x]
p1 = ListPlot[data]
p2 = Plot[fun, {x, xmin, xmax}]
Show[p1, p2]
```

**OPGAVE 5.2** Deze opgave kan het beste met MATHEMATICA worden uitgevoerd. (Het kan ook met andere rekenprogramma's zoals MATHCAD of MAPLE of met een spreadsheet. Of schrijf zelf een Pascal, C, Fortran of Basic programma!)

De  $\Delta G$  van een reactie werd uit de evenwichtsconstante bij verschillende temperaturen bepaald. De volgende waarden werden gevonden:

$T/\text{K}$	$\Delta G/\text{kJ mol}^{-1}$
270	40,3
280	38,2
290	36,1
300	32,2
310	29,1
320	28,0
330	25,3

De fout in  $T$  is verwaarloosbaar en de gewichten van alle bepalingen zijn gelijk. Bepaal nu de reactie-entropie  $\Delta S = -d\Delta G/dT$  door de waarden van  $\Delta G$  aan een lineaire functie van  $T$  aan te passen. Geef ook de nauwkeurigheid van  $\Delta S$ . Extrapoler  $\Delta G$  bij  $T = 350$  K en geef de standaardafwijking die volgt uit de variantie en covariantie van de parameters van de kleinste kwadratenaanpassing.

Doe nu hetzelfde, maar neem voor de  $x$ -waarden de temperatuur  $T - 300$  in plaats van  $T$  zelf. Discussieer de verschillen tussen de twee berekeningen.

(voor antwoorden zie pag. 105)

## 5.2 De algemene kleinste-kwadratenaanpassing

Wanneer de aan te passen functie geen lineaire functie van de parameters is, dan is het nog steeds mogelijk om een kleinste kwadratenaanpassing te maken. Soms is het mogelijk om de functie te lineariseren (zie par. 4.2 op pag. 47). Zo is bijvoorbeeld de functie  $f(t) = a \exp(-kt)$  niet lineair in de parameter  $k$ . Maar door de logaritme te nemen:

$$\ln f(t) = -kt + \ln a \quad (5.18)$$

krijgen we een functie die wel lineair in  $k$  is en de vorm  $ax + b$  heeft. Dus kan de normale lineaire regressie op de meetpunten  $(t_i, \ln y_i)$  worden toegepast. Maar denk wel aan de gewichtsfactoren: als alle punten  $y$  gelijke standaarddeviaties  $\sigma$  hebben, dan hebben de waarden  $\ln y_i$  dat niet:

$$\sigma_{\ln y} = \left| \frac{d \ln y}{dy} \right| \sigma = \sigma/y_i \quad (5.19)$$

In dit geval moet men voor de gewichtsfactoren  $w_i$  de waarde  $1/y_i^2$  nemen of iets dat daar mee evenredig is. Negatieve meetwaarde van  $y$ ,

die door toevallige afwijkingen bij grote waarden van  $t$  voor kunnen komen, kunnen niet verwerkt worden. Zij mogen niet selektief weggelaten worden omdat dan de wèl verwerkte positieve waarden het resultaat eenzijdig beïnvloeden. Men kan het beste alle meetpunten weglaten boven de waarde van  $t$  waar voor het eerst een negatieve waarde optreedt. Fraai is dit alles niet.

Vaak is zo'n linearisatie niet alleen niet fraai, maar ook niet mogelijk. Bijvoorbeeld de functie

$$f(t) = a \exp(-kt) + b \quad (5.20)$$

kan niet tot een lineaire functie van de parameters  $a$ ,  $b$ , en  $k$  herschreven worden. Dan moet een *niet-lineaire kleinste kwadratenprocedure* van stal gehaald worden. Er zijn meerdere computerprogramma's in standaard-programmabibliotheken die hiervoor geschikt zijn. Zij geven ook vaak de covariantiematrix die voor het vaststellen van de nauwkeurigheden in de gevonden parameters gebruikt kunnen worden. Hieronder is aangegeven hoe een niet-lineaire kleinste kwadratenaanpassing met MATHEMATICA kan worden uitgevoerd.

### MATHEMATICA VOORBEELD 5.3

Hieronder volgt een kort programma `bestfit` dat twee parameters  $a$  en  $b$  in een functie `fun[a_,b_,x_]` aanpast, zodat een gegeven lijst van 'meetpunten' `ylist` een kleinste-kwadratenafwijking heeft van de functiewaarden voor een gegeven lijst van  $x$ -waarden `xlist`. Een lijst `siglist` van standaarddeviaties van de meetpunten moet gegeven worden. Ook zijn beginschattingen van  $a$  en  $b$ : `astart` en `bstart` vereist. Het programma produceert de beste waarden van  $a$  en  $b$ : `amin` en `bmin` en de minimale som van de gewogen kwadratische afwijkingen `chisqmin`. De functie `chisq[a,b]` kan gebruikt worden om nauwkeurigheden te schatten (zie par. 5.4)

```
bestfit := (
  diflist[a_,b_] = ylist - fun[a,b,xlist];
  wlist = 1.0/siglistverb'^2;
  chisq[a_,b_] = Apply[Plus, diflist[a,b]^2*wlist];
  result = FindMinimum[chisq[a,b],{a,astart},{b,bstart}];
  amin = result[[2]][[1]][[2]];
  bmin = result[[2]][[2]][[2]];
  chisqmin = result[[1]];
  Print["amin=",amin," bmin=",bmin," chisqmin=",chisqmin]
)
```

We passen dit als voorbeeld toe op de enzymkinetiek-gegevens van Tabel 4.2 (pag. 51). Daarbij bevat `xlist` de substraatconcentraties  $[S]$ , `ylist` de gemeten snelheden  $v$ , `siglist` de s.d.'s in  $v$ , `a` de parameter  $v_{max}$ , `b` de parameter  $K_m$  en `astart` en `bstart` de beginwaarden die in opgave 4.2 zijn bepaald.

```
xlist = {30, 60, 100, 150, 250, 400}
ylist = {3.09, 5.52, 7.59, 8.72, 10.69, 12.34}
siglist = Table[0.2,{i,6}]
fun[a_,b_,x_] = a*x/(b+x)
astart = 15.
bstart = 105.
bestfit
```

De uitkomst is:  $amin = 15,7521$ ;  $bmin = 114,648$ ;  $chisqmin = 4,27538$ . Hiermee hebben we wel de beste aanpassing berekend, maar we hebben nog geen idee van de nauwkeurigheid van het antwoord.

Niet-lineaire kleinste-kwadratenprogramma's hebben in het algemeen beginschattingen van de parameters nodig; als deze er ver naast zitten, kunnen er onzinnige of helemaal geen antwoorden uit de computer komen. Het beste kunnen deze beginschattingen uit een grove grafische analyse worden afgeleid, zoals in par. 4.3 is behandeld.

Met het bepalen van de beste parameters in de functie zijn we er nog niet! We zouden graag een idee hebben of de aanpassing geslaagd is en we willen ook betrouwbaarheidsschattingen hebben voor de gevonden parameters. De sleutel voor het beantwoorden van deze vragen ligt besloten in de waarde van de som van kwadraten in het minimum en het gedrag van deze som als functie van de parameters. Hoe dat zit zien we in de volgende twee paragrafen.

**OPGAVE 5.3** Maak een kleinste-kwadratenaanpassing van de vier-parameterfunctie  $a \exp(-px) + b \exp(-qx)$  aan de meetwaarden  $x, y$  uit tabel 4.1 op pag. 48. Modificeer hiervoor het bovenvermelde MATHEMATICA voorbeeld `bestfit`. Gebruik de beginschattingen die in par. 4.2 op grafische wijze zijn bepaald. Als het programma geen resultaat bereikt na het maximaal toegestane aantal iteraties, neem dan de eindwaarden van de parameters als nieuwe startwaarden en pas `bestfit` nogmaals toe.

### 5.3 De chi-kwadraattest

Stel dat we uit een serie meetgegevens een functieaanpassing hebben gemaakt met de kleinste-kwadratenmethode. Het is dan nuttig om te beoordelen of de aanpassing gelukt is: voldoen de metingen binnen de meetfout aan het functionele verband dat we hebben verondersteld?

We kunnen naar beide zijden afwijkingen verwachten: heeft de functie te weinig parameters of een verkeerde vorm dan zullen de meetpunten systematische afwijkingen van de functie vertonen die groter zijn dan de toevallige fouten; gebruiken we te veel parameters in de functie, dan zal (als de aanpassing al lukt) de functie ook proberen de toevallige afwijkingen te volgen en worden de afwijkingen juist te klein.

Het eerste dat gedaan moet worden is de meetwaarden met hun standaarddeviatie en de berekende functie tegelijk in een grafiek uit te zetten en te kijken of er systematische afwijkingen zijn. Is dat op het oog niet het geval, dan kan een *chi-kwadraattest*<sup>1</sup> worden uitgevoerd. Deze stelt vast wat de waarschijnlijkheid is dat de gemeten som van kwadratische afwijkingen het gevolg is van toevallige meetfouten in de punten. *Om de test zinvol te kunnen uitvoeren is het nodig goede schattingen te hebben van de standaarddeviaties  $\sigma_i$  van elk meetpunt.* Hoe we te werk moeten gaan als we deze kennis niet hebben, wordt aan het eind van deze paragraaf besproken.

Hier toe wordt eerst  $\chi^2$  bepaald. Dit is de waarde van de minimale som van kwadratische afwijkingen als voor de gewichten  $w_i$  de waarden van  $1/\sigma_i^2$  zijn gebruikt:

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - f_i)^2}{\sigma_i^2} \quad (5.21)$$

Elk van de termen in deze som zou gemiddeld de waarde 1 moeten hebben en de som zelf zou dus ongeveer  $n$  moeten bedragen. Helemaal juist is dat niet, want bij een lineaire regressie op  $f = ax + b$  zijn er twee vrijheidsgraden ‘verbruikt’ voor het vaststellen van twee parameters: het aantal vrijheidsgraden  $\nu$  bedraagt dan  $n - 2$  en  $\chi^2$  zal ongeveer gelijk zijn aan  $\nu = n - 2$ . In het algemeen, als er  $m$  parameters aangepast zijn, dan bedraagt het aantal vrijheidsgraden nog  $\nu = n - m$ . Maar  $\chi^2$  heeft natuurlijk een waarschijnlijkheidsverdeling rond die waarde. De verdelingsfunctie van  $\chi^2$  hangt ook af van het aantal vrijheidsgraden  $\nu$  en wordt met  $f(\chi^2|\nu)$  aangeduid. De functie  $f(\chi^2|\nu)$  is vrij ingewikkeld

<sup>1</sup>Spreek de ch uit als de G in Groningen

(zie het blad CHIKWADRAATVERDELING op pag. 115 voor formules), maar heeft gelukkig de eigenschap dat hij voor grote  $\nu$  de vorm van een normale verdeling krijgt. Het gemiddelde van de verdeling is gelijk aan  $\nu$  en de standaarddeviatie is gelijk aan  $\sqrt{2\nu}$ .

Tabellen van de  $\chi^2$ -verdeling geven niet de verdelingsfunctie van  $\chi^2$  zelf, maar de *cumulatieve verdelingsfunctie*  $F(\chi^2|\nu)$ : de waarschijnlijkheid dat de gevonden waarde van chi-kwadraat kleiner is dan  $\chi^2$ . De functie  $1 - F(\chi^2|\nu)$  is dan vanzelfsprekend de kans dat de waarde  $\chi^2$  overschreden wordt. Het blad CHIKWADRAATVERDELING geeft ook een tabel van de waarden van  $\chi^2$  die nog acceptabel zijn bij acceptatiegrenzen van 1%, 10%, 50%, 90% en 99%. In de meeste statistiekboeken (en in het *Handbook of Chemistry and Physics*) vindt men uitgebreidere tabellen.

Het gebruik van de tabel is als volgt. Van te voren moet men een acceptatie-eis stellen, bijv.: tussen 1% en 99%, of tussen 10 en 90%. Deze keuze is subjectief en de gekozen acceptatiegrenzen moeten bij publicatie worden vermeld. Voor ons voorbeeld kiezen we de 10-90% grenzen. Valt een  $\chi^2$ -waarde onder de 10%-waarde, dan wordt de afwijking niet meer als toevallig geaccepteerd en is de conclusie dat de aanpassing *te goed* is. De aangepaste functie heeft dan te veel parameters en de aanpassing moet nog eens over gedaan worden met een eenvoudiger functie. *De meetgegevens rechtvaardigen de complexiteit van de functie niet.* Valt de  $\chi^2$ -waarde echter boven het 90%-niveau, dan wijken de gegevens significant van de functie af. In dat geval is de gebruikte functie ook ongeschikt om de data te beschrijven en moet naar een andere functie worden gezocht, mogelijk met meer parameters.

Het volgende voorbeeld moge dit verduidelijken. Stel we hebben 10 meetpunten  $(x, y)$ , waarbij  $x$  nauwkeurig is en  $y$  een bekende standaarddeviatie heeft. Een eenvoudige theorie voorspelt dat  $y$  een lineaire functie van  $x$  is  $(ax + b)$ , maar een verfijnde theorie voorspelt een tweedegraads functie  $px^2 + qx + r$ . Rechtvaardigen de gegevens de verfijnde theorie boven de eenvoudige theorie met een betrouwbaarheidsniveau van 90%? Hiertoe voeren we een kleinste kwadratenaanpassing op beide functies uit, gebruik makend van  $1/\sigma_i^2$  als gewichtsfactor voor het  $i^e$  meetpunt. Voor de lineaire functie vinden we  $\chi^2 = 14,2$  en voor de kwadratische functie  $\chi^2 = 7,3$ . Kijken we nu in de tabel voor 8 vrijheidsgraden, dan zien we dat 14,2 boven de 90% grens ligt en dus niet acceptabel is volgens ons criterium. De kwadratische functie (er zijn nu 7 vrijheidsgraden) is wel acceptabel. Hadden we bijvoorbeeld de waarden 12,3 en 6,5 gevonden, dan hadden we hieruit niet kunnen besluiten dat de tweedegraads func-



tie noodzakelijk was, ondanks het feit dat die een veel betere aanpassing geeft dan de lineaire functie.

Wat te doen als we de standaarddeviaties van de meetpunten niet goed kennen? De chi-kwadraatanalyse is dan niet bruikbaar. Het kan bijvoorbeeld goed zijn dat we de meetfouten te groot hebben ingeschat; de waarde voor  $\chi^2$  komt dan te klein uit en we zouden ten onrechte besluiten dat de functieaanpassing te nauwkeurig is<sup>2</sup>. Hebben we de meetfouten te klein ingeschat dan komt  $\chi^2$  te groot uit en zouden we ten onrechte naar een functie zoeken die een betere aanpassing geeft. Het effect van een verkeerde inschatting is vrij groot: hebben we bijv.  $\sigma$  een factor 2 te groot ingeschat, dan wordt  $\chi^2$  een factor 4 kleiner. Vinden we dan bij 10 vrijheidsgraden een waarde van 2, dan valt dit onder de 1% grens (zie de tabel), terwijl bij juiste inschatting van  $\sigma$  de perfect aanvaardbare waarde 8 verkregen zou zijn.

Het lijkt erop dat we bij ontbreken van een goede schatting van de standaarddeviaties niets met de gevonden  $\chi^2$  waarde kunnen beginnen, maar dat is onjuist. Hebben een voldoende groot aantal metingen dan kunnen we de gevonden  $\chi^2$  waarde juist gebruiken om de  $\sigma$  te bepalen. We kunnen dan alleen een globale schaling van de  $\sigma$ 's vinden, maar vinden natuurlijk geen waarden voor de individuele weegfactoren. Het recept is eenvoudig: schaal alle  $\sigma$ 's met de wortel uit het quotiënt van de gemeten waarde van  $\chi^2$  en de waarde  $\nu = n - m$  die  $\chi^2$  gemiddeld zou moeten hebben bij het gegeven aantal vrijheidsgraden. Bijvoorbeeld: vind je bij 50 meetpunten en 2 parameters een  $\chi^2$  van 20, vermenigvuldig de vooraf geschatte waarden van  $\sigma_i$  dan met  $\sqrt{20/48}$ . Dit gaat zelfs goed als de geschatte  $\sigma$ 's onzinnig waren, omdat zij bijvoorbeeld bij gebrek aan enige kennis gelijk aan 1 waren genomen.

Het is nu natuurlijk niet meer mogelijk  $\chi^2$  te gebruiken om de kwaliteit van de aanpassing te beoordelen. Het is raadzaam daarvoor andere criteria toe te passen, via een analyse van de *residuen*  $\epsilon_i = y_i - f_i$ . Deze moeten een *random* karakter hebben. Als je ze in een grafiek uitzet, mag er geen systematisch verloop in zitten. Ook de (cumulatieve) verdelingsfunctie moet een min of meer Gaussisch karakter hebben. Soms wordt

---

<sup>2</sup>Hetzelfde gebeurt als we gesjoemeld hebben met de metingen, bijvoorbeeld metingen die ons niet bevallen er uit gegooid hebben, maar dit gedrag behoort tot de (helaas veel voorkomende) wetenschappelijke fraude.

de *autocorrelatiefunctie*  $C(k)$  bepaald:

$$C(k) = \frac{1}{n-k} \sum_{i=1}^{n-k} \epsilon_i \epsilon_{i+k}; \quad (5.22)$$

deze mag geen ‘structuur’ vertonen.

**OPGAVE 5.4** Ga met behulp van de tabel op p. 2 van het blad CHIKWADRAATVERDELING na of de waarde van  $\chi^2$  die gevonden is bij de kleinste-kwadratenaanpassing van de ureasekinetiek (pag. 67) acceptabel is. Hoeveel vrijheidsgraden zijn er? Bereken ook met MATHEMATICA hoe groot de cumulatieve distributiefunctie van de  $\chi^2$ -verdeling is bij de gevonden waarde van  $\chi^2$  (zie MATHEMATICA voorbeeld op pag. 20).

## 5.4 Nauwkeurigheid van de parameters

Veronderstel dat de residuen onze tests hebben doorstaan en dat we de waarden van  $\chi^2$  kunnen vertrouwen: òf we kenden de standaarddeviaties van de meetpunten nauwkeurig, òf we hebben gebruik gemaakt van de gevonden waarde van de minimale som van kwadratische afwijkingen om de s.d.’s te schalen, zodat  $\chi^2$  in het minimum nu precies gelijk is aan  $n - m$ . Nu kunnen we de varianties en covarianties van de parameters berekenen.

We geven hier enkele algemene formules, die in Bijlage H worden afgeleid. We gaan weer uit van  $n$  meetpunten  $y_i$  en veronderstellen dat we een (al of niet lineaire) functie van  $m$  parameters  $\theta_k, k = 1, \dots, m$  willen aanpassen. De kleinste-kwadratenprocedure geeft ons  $\chi^2$  als functie van de parameters, met een minimum  $\chi_0^2$  voor de parameterwaarden  $\hat{\theta}_i$ , die als beste schattingen kunnen worden beschouwd. Aangezien  $\chi_0^2$  een minimum is, is de functie  $\chi^2(\theta_1, \theta_2, \dots, \theta_m)$  *kwadratisch* in de buurt van het minimum (de kwadratische term is de eerste term in een Taylorontwikkeling van  $\chi^2$  rond het minimum, ook voor een aan te passen functie die niet lineair is in de parameters).

Een belangrijke rol in de foutendiscussie speelt de matrix  $\mathbf{B}$  met elementen

$$B_{kl} = \sum_{i=1}^n \frac{1}{\sigma_i^2} \frac{\partial f_i}{\partial \theta_k} \frac{\partial f_i}{\partial \theta_l} \quad (5.23)$$

De partiële afgeleiden van  $f$  naar  $\theta$  zijn constanten voor lineaire functies; voor niet-lineaire functies moeten hiervoor de afgeleiden op de beste-fit waarden  $\hat{\theta}$  worden genomen. Deze matrix is tevens de (helft van de) matrix van tweede afgeleiden van  $\chi^2$  (zie Bijlage H):

$$B_{kl} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta_k \partial \theta_l}, \quad (5.24)$$

zodat  $\mathbf{B}$  de kromming van de functie  $\chi^2(\theta_1, \theta_2, \dots, \theta_m)$  rond het minimum bepaalt:

$$\Delta \chi^2 = \chi^2(\theta) - \chi^2(\hat{\theta}) \quad (5.25)$$

$$= \sum_{k,l=1}^m B_{kl} \Delta \theta_k \Delta \theta_l \quad (5.26)$$

De covarianties van de parameters zijn nu gegeven door de *inverse* van de matrix  $\mathbf{B}$  (zie Bijlage H). Geven we die met  $\mathbf{C}$  aan:

$$\mathbf{C} = \mathbf{B}^{-1} \quad (5.27)$$

dan is

$$\text{cov}(\theta_k, \theta_l) = C_{kl}. \quad (5.28)$$

$\mathbf{C}$  is dus de covariantiematrix van de parameters. Zo is de bijvoorbeeld de standaarddeviatie van  $\theta_k$  gelijk aan

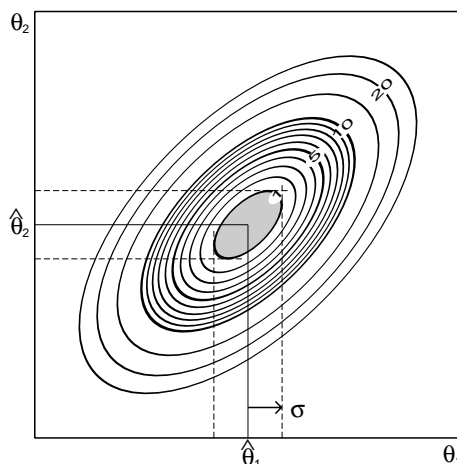
$$\sigma_{\theta_k} = \sqrt{C_{kk}} \quad (5.29)$$

en de correlatiecoëfficiënt tussen  $\theta_k$  en  $\theta_l$  is

$$\rho_{kl} = \frac{C_{kl}}{\sqrt{C_{kk} C_{ll}}}. \quad (5.30)$$

Kennis van de matrix  $\mathbf{C}$  is dus voldoende om de nauwkeurigheden in de parameters en hun onderlinge correlaties af te schatten.

Het is ook mogelijk de standaarddeviaties af te lezen van een plot van  $\chi^2$  tegen de parameters. Voor twee parameters wordt een paraboloïde verkregen, zoals in figuur 5.2 als een contourplot is uitgezet. De hoogtelijnen geven waarden voor  $\Delta \chi^2$ , d.w.z. de waarde boven het minimum. Men kan nu aantonen (zie Bijlage H) dat de *raakklijnen* aan de contour voor  $\Delta \chi^2 = 1$  de grenzen van  $\hat{\theta} \pm \sigma_{\theta}$  aangeven.



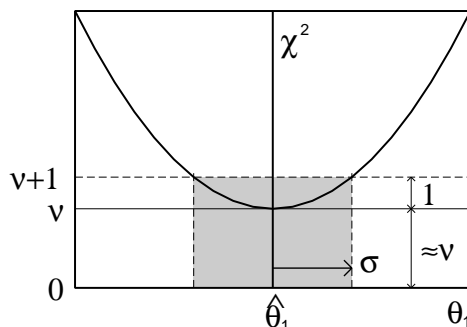
Figuur 5.2: Een contourplot van  $\Delta\chi^2$  als functie van twee parameters  $\theta_1$  en  $\theta_2$ . De getekende raaklijnen aan de ellips voor  $\Delta\chi^2 = 1$  geven de standaarddeviaties van  $\theta_1$ , resp.  $\theta_2$ .

Zijn er meer dan twee parameters, dan kan zo'n contourplot nog steeds voor twee geselecteerde parameters worden gemaakt, maar dan moet wel voor elk punt in de plot de waarde van  $\chi^2$  geminimaliseerd zijn t.o.v. alle andere parameters. Zo kan ook een ééndimensionale plot voor één parameter  $\theta_1$  worden gemaakt, waarbij voor elke waarde van  $\theta_1$  de functie  $\chi^2$  geminimaliseerd is t.o.v. alle andere parameters. Zo'n curve is een parabool (fig. 5.3) waaruit de s.d. van  $\theta_1$  kan worden afgelezen. De ellipsen (twee dimensies) zijn *projecties* van de meerdimensionale ellipsoiden die  $\chi^2$  als functie van alle parameters voorstellen.

Het oppervlak binnen elke ellips in fig. 5.2 komt overeen met parameterwaarden die binnen een gegeven *acceptatieniveau* vallen. Zo heeft de ellips bij  $\Delta\chi^2 = 2,30$  een acceptatieniveau van 68,3%, de ellips bij  $\Delta\chi^2 = 4,61$  een acceptatieniveau van 90% en de ellips bij  $\Delta\chi^2 = 9,21$  een acceptatieniveau van 99%. Deze waarden gelden voor een ellips in twee dimensies. We gaan hier verder niet op in en verwijzen de geïnteresseerde lezer naar *Numerical Recipes* [8] (zie literatuurlijst op pag. 77).

#### MATHEMATICA VOORBEELD 5.4

Maken we voor het voorbeeld van de ureasekinetiek (pag. 67) een contourplot van  $\chi^2$  als functie van  $a$  en  $b$  dan verkrijgen we figuur



Figuur 5.3: De waarde van  $\chi^2$  uitgezet tegen één parameter, maar geminimaliseerd t.o.v. alle andere parameters. Uit de grenzen bij  $\Delta\chi^2 = 1$  volgt de standaarddeviatie van  $\hat{\theta}_1$ .

5.4. Hier zijn contouren uitgezet voor  $\Delta\chi^2 = 1$  (de binnenste contour, bedoeld om de standaarddeviaties van  $a = v_{max}$  en  $b = K_m$  af te lezen),  $\chi^2 = 2,30$  (68% acceptatie),  $\chi^2 = 4,61$  (90% acceptatie) en  $\chi^2 = 9,21$  (99% acceptatie). Deze plot is gemaakt met

```
ContourPlot[chisq[a,b],{a,amin-2.,amin+2.},
  {b,bmin-30.,bmin+30.},ColorFunction -> Hue,
  Contours -> {4.3,5.3,6.6,8.9,13.5}]
```

Merk op dat MATHEMATICA geen erg gladde contouren aflevert. Lezen we de projecties van de binnenste ellips op de assen af dan vinden we ongeveer voor  $\sigma_a = 0,4$  en  $\sigma_b = 7$ , zodat de gevonden parameters zijn:

$$v_{max} = 15,8 \pm 0,4 \text{ mmol min}^{-1} \text{ mg}^{-1}$$

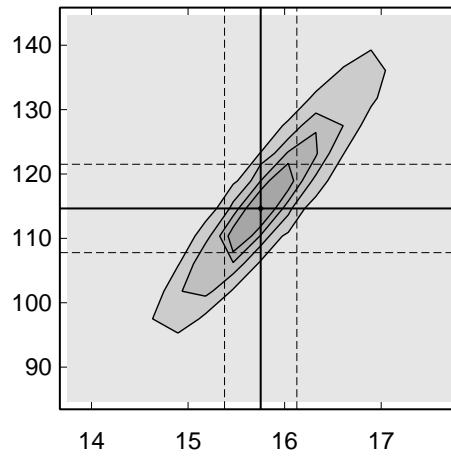
$$K_m = 115 \pm 7 \text{ mM.}$$

De correlatiematrix  $C$  kan ook berekend worden en geeft een correlatiecoëfficiënt van 0,93 tussen  $v_{max}$  en  $K_m$ .

#### MATHEMATICA VOORBEELD 5.5

Het voorbeeld van de ureasekinetiek kunnen we ook gebruiken om de covariantiematrix van de parameters via formule 5.28 te berekenen. Daarvoor moeten we dus de matrix  $\mathbf{B}$  (vgl. 5.24) uitrekenen. We gaan weer uit van de procedure `bestfit` op pag. 67. Voeg hier, nadat `bestfit` is uitgevoerd, de volgende regels aan toe:

```
dfda[a_,b_,x_] = D[fun[a,b,x],a];
dfdb[a_,b_,x_] = D[fun[a,b,x],b];
aa := dfda[amin,bmin,xlist];
```



Figuur 5.4: Een contourplot van  $\chi^2$  voor de ureasekinetiek, uitgezet tegen  $a = v_{max}$  (horizontaal) en  $b = K_m$  (vertikaal). De contouren hebben van binnen naar buiten de waarden 1,00, 2,30 (68% acceptatie), 4,61 (90%) en 9,21 (99%) voor  $\Delta\chi^2$  (zie tekst).

```
bb := dfdb[amin,bmin,xlist];
bmat = {{wlist.(aa aa), wlist.(aa bb)},
        {wlist.(bb aa), wlist.(bb bb)}};
cmat = Inverse[bmat]
```

De eerste twee regels definiëren de (symbolische) afgeleiden van de functie naar  $a$  en  $b$ ; de derde en vierde regels geven alleen korte namen; de vijfde regel rekent de matrix  $\mathbf{B}$  uit en de zesde regel invertteert die. De laatste uitkomst is dus de matrix van covarianties van  $a$  en  $b$ . Hier kunnen de s.d. en correlatiecoëfficiënt uitgehaald worden:

```
sig_a = Sqrt[cmat[[1,1]]]
sig_b = Sqrt[cmat[[2,2]]]
rho_ab = cmat[[1,2]]/sig_a/sig_b
```

Als we dit uitvoeren krijgen we:  $\sigma_a = 0,400$ ;  $\sigma_b = 7,37$ ;  $\rho_{ab} = 0,927$ . Vergelijk dit met de waarden van  $\sigma_a$  en  $\sigma_b$  verkregen uit figuur 5.4. We zien dat er een grote correlatie tussen  $a$  en  $b$  is.



# Bibliografie

- [1] P. R. Bevington and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, 3rd ed., McGraw-Hill 2003 (oorspr. 1969).  
*Dit boek (met Fortran programma's) is geschreven voor fysici. De onderwerpen komen redelijk overeen met die in het voor u liggende boekje. Wel enigszins verouderd, maar geschikt om voor verdere studie te gebruiken.*
- [2] B. P. Roe, *Probability and Statistics in Experimental Physics*, Springer-Verlag 1992.  
*Gericht op (nucleaire) fysica en ietsje meer highbrow dan [1].*
- [3] R. V. Hogg and J. Ledolter, *Applied Statistics for Engineers and Physical Scientists*, 2nd ed., Macmillan Publ. Cy 1992 (1st ed. 1987).  
*Dit boek is meer gericht op industriële kwaliteitscontrole en 'experimental design' dan op foutenanalyse van experimentele resultaten. Het omvat wel de nodige basisstatistiek.*
- [4] R. E. Walpole, R. H. Myers en S. L. Myers, *Probability and Statistics for Engineers and Scientists*, 6th ed., Prentice Hall 1998.  
*Een goed bruikbare 'klassieker' (eerste editie 1972), omvat veel methoden en toepassingen. Wiskundig toegankelijk.*
- [5] J. D. Petroccelli, B. Nandram en M. Chen, *Applied Statistics for Engineers and Scientists*, Prentice Hall 1999.  
*Een uitgebreid boek over moderne statistische methoden met veel praktische voorbeelden. De wiskundige achtergrond wordt niet uitgewerkt en er zijn weinig referenties naar de literatuur.*



- [6] R. Barlow, *Statistics - A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, New York, 1989.  
*Eenvoudig; toegespitst op de natuurwetenschappen.*
- [7] J. C. Miller en J. N. Miller, *Statistics for Analytical Chemistry*, 3rd ed., Ellis Horwood PTR Prentice Hall, New York, 1993.  
*Statistiek toegepast op de analytische chemie. Geeft veel praktische methoden, maar geeft geen enkele achtergrond.*
- [8] W. H. Press, A. A. Teukolsky, W. T. Vetterling en B. P. Flannery, *Numerical Recipes, The Art of Scientific Computing*, 2nd ed., Cambridge University Press 1992.  
*Een uitgebreide en praktische handleiding compleet met computer-programma's, maar niet altijd gemakkelijk. Bevat hoofdstukken over statistiek en verwerking van meetgegevens.*
- [9] M. Abramowitz en I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.  
*Een uitgebreid tabellenboek met mathematische functies, waaronder vele statistische functies. Geen leerboek over statistiek.*
- [10] W. H. Beyer, *CRC Standard Probability and Statistics Tables and Formulae*, CRC Press, Boca Raton, Florida, 1991.  
*Zeer uitgebreid tabellenboek, zonder veel uitleg. Voorkennis van statistiek is nodig voor gebruik.*
- [11] *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, Florida, elk jaar.  
*Bevat ook statistische tabellen.*
- [12] G. E. P. Box en G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass, 1973.  
*Uitgebreide en geavanceerde behandeling van statistische verwerking, gebaseerd op de voorspelling van de waarschijnlijkheidsverdeling van de te bepalen grootheden.*
- [13] D. Birkes en Y. Dodge, *Alternative Methods of Regression*, Wiley, New York 1993. *Behandelt lineaire regressie met verschillende methoden en criteria. Vrij gespecialiseerd boek.*

## Bijlage A

# Optellen van fouten

*Waarom tellen fouten kwadratisch op?*

Stel dat we de som  $f = x + y$  bepalen van twee grootheden, die ieder getrokken zijn uit een kansverdeling, met

$$E[x] = \mu_x; \quad E[(x - \mu_x)^2] = \sigma_x^2 \quad (\text{A.1})$$

$$E[y] = \mu_y; \quad E[(y - \mu_y)^2] = \sigma_y^2 \quad (\text{A.2})$$

De grootheid  $f = x + y$  heeft een verwachting

$$\mu = \mu_x + \mu_y \quad (\text{A.3})$$

en een variantie

$$\begin{aligned} \sigma_f^2 &= E[(f - \mu)^2] = E[(x - \mu_x + y - \mu_y)^2] = \\ &= E[(x - \mu_x)^2 + (y - \mu_y)^2 + 2(x - \mu_x)(y - \mu_y)] = \\ &= \sigma_x^2 + \sigma_y^2 + 2E[(x - \mu_x)(y - \mu_y)] \end{aligned} \quad (\text{A.4})$$

Als  $x$  en  $y$  onafhankelijk van elkaar zijn (d.w.z. dat de afwijkingen van het gemiddelde van  $x$  en  $y$  statistisch onafhankelijke samples zijn), dan is de laatste term nul en tellen de varianties inderdaad kwadratisch op.

We zien echter tegelijk aan deze afleiding dat fouten niet meer kwadratisch optellen wanneer de twee grootheden afhankelijk zijn. De grootheid  $E[(x - \mu_x)(y - \mu_y)]$  heet de *covariantie* van  $x$  en  $y$ , die ook wel als *correlatiecoëfficiënt*  $\rho_{xy}$  wordt gegeven:

$$\text{cov}(x, y) = E[(x - \mu_x)(y - \mu_y)] \quad (\text{A.5})$$

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}. \quad (\text{A.6})$$

De complete formule is dus eigenlijk

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2 \text{cov}(x, y). \quad (\text{A.7})$$

Voor een *verschil*  $f = x - y$  geldt

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2 \text{cov}(x, y). \quad (\text{A.8})$$

Voor een product, resp. quotiënt, gelden dezelfde formules voor de *relatieve* deviaties:

$$\frac{\text{var } f}{f^2} = \frac{\text{var } x}{x^2} + \frac{\text{var } y}{y^2} \pm 2 \frac{\text{cov}(x, y)}{xy}, \quad (\text{A.9})$$

waarbij het plusteken geldt voor  $f = xy$  en het minteken voor  $f = x/y$ .

De algemene formule voor de variantie van een functie  $f(x_1, x_2, \dots)$  is

$$\text{var}(f) = \sum_{i,j} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \text{cov}(x_i, x_j) \quad (\text{A.10})$$

Hier is  $\text{cov}(x_i, x_i) = \text{var}(x_i)$ . Deze formule volgt direct uit het kwadrateren van

$$df = \sum_i \frac{\partial f}{\partial x_i} dx_i.$$

Hieronder volgt een voorbeeld van het gebruik van covarianties. Stel we hebben (met de computer) een kleinste-kwadratenaanpassing gedaan van  $f(x) = ax + b$  op een aantal meetpunten met het resultaat:

$$a = 2,30526; \quad b = 5,21632$$

$$\sigma_a = 0,00312; \quad \sigma_b = 0,0357; \quad \rho_{ab} = 0,87326$$

Deze aanpassing willen we gebruiken voor een intra- of extrapolatie: *welke waarde en s.d. wordt verwacht voor  $f(10)$ ?*

Hiervoor rekenen we eerst de waarden, varianties en covarianties uit voor de twee grootheden  $ax$  en  $b$ , die we willen optellen.  $x$  is hierbij een vermenigvuldigingsfactor, die kwadratisch in  $\text{var}(ax)$  en lineair in  $\text{cov}(ax, b)$  terecht komt:

$$ax = 23,0526; \quad b = 5,21632; \quad f = 28,26892$$

$$\text{var}(ax) = 0,00312^2 \times 10^2; \quad \text{var}(b) = 0,0357^2$$

$$\text{cov}(ax, b) = 10 \times 0,87326 \times 0,00312 \times 0,0357$$

Invullen in formule A.7 geeft dan  $\text{var}(f) = 0,00419$ . Hadden we de covariantie niet meegeteld, dan was  $\text{var}(f) = 0,00225$  geweest. De s.d. van  $f$  is nu 0,0648 en we geven het resultaat, naar boven afgerond, weer als  $f = 28,27 \pm 0,07$ .



## Bijlage B

# Systematische fouten in doorwerking

Wanneer  $f(x)$  een niet te verwaarlozen kromming heeft, dan kunnen er systematische fouten in  $f$  optreden, die een gevolg zijn van toevallige fouten in  $x$ , ook als deze zelf symmetrisch verdeeld zijn. Stel we hebben een verzameling bolletjes die ongeveer, maar niet precies, dezelfde straal hebben. We meten de straal van een aantal bolletjes en vinden ongeveer een normale verdeling met  $r = 1,0 \pm 0,1$  mm. Dit levert voor het volume van de bol op (in te veel decimalen):  $V = \frac{4}{3}\pi r^3 = 4,19$  mm<sup>3</sup>. Werken we de derde macht echter uit:

$$(r \pm \Delta r)^3 = r^3 \pm 3r^2\Delta r + 3r(\Delta r)^2 \pm (\Delta r)^3$$

en gaan we er van uit dat de verdelingsfunctie voor  $\Delta r$  symmetrisch is, dan zien we dat de derde term gemiddeld niet nul is en dus een bijdrage zal geven aan de te verwachten waarde voor  $f$ :

$$E[r^3] = r^3 + 3r \text{ var}(r)$$

Als  $E[f(x)] \neq f(E[x])$  spreken we van een systematische afwijking of een ‘bias’. Voor ons voorbeeld bedraagt deze extra bijdrage aan het volume 0,13 mm<sup>3</sup> en het verwachte volume is dus 4,32 mm<sup>3</sup>. Maken we deze correctie niet, dan heeft het voorspelde volume een ‘bias’ van -0,13. Dit is weliswaar 10× kleiner dan de s.d., maar soms moet men met deze ‘bias’ rekening houden.

De algemene formule komt uit de tweede term van een Taylorreeks:

$$f(x) = f(a) + (x - a)f'(a) + \frac{1}{2}(x - a)^2 f''(a) + \dots \quad (\text{B.1})$$

$$E[f] = f(E[x]) + \frac{1}{2} \frac{d^2 f}{dx^2} \text{var}(x) + \dots \quad (\text{B.2})$$

## Bijlage C

# De binomiaal- en de Poissonverdeling

Als de uitkomst van een steekproef 0 of 1 kan zijn, en de kans op 1 is  $p$ , dan zijn er voor twee onafhankelijke waarnemingen de volgende mogelijkheden: 00, 01, 10, 11. De kans  $f(k)$  dat precies  $k$  keer een 1 wordt waargenomen is

$$\begin{aligned}f(0) &= (1-p)^2 \\f(1) &= 2p(1-p) \\f(2) &= p^2\end{aligned}$$

In het algemeen is de kans  $f(k;n)$  dat uit  $n$  onafhankelijke metingen precies  $k$  keer een 1 wordt waargenomen, gelijk aan

$$f(k;n) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad (\text{C.1})$$

waar

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (\text{C.2})$$

Dit is de *binomiaalverdeling*.

Zijn er meer dan twee (bijv.  $m$ ) mogelijke uitkomsten (met waarschijnlijkheden  $p_1, p_2, \dots, p_m$ , dan ontstaat de *multinomiaalverdeling*:



$$f(k_1, k_2, \dots, k_m; n) = \frac{n!}{k_1! k_2! \dots k_m!} \prod_{i=1}^m p_i^{k_i} \quad (\text{C.3})$$

De binomiaalverdeling is (uiteraard) genormeerd:

$$\sum_{k=0}^n f(k; n) = 1 \quad (\text{C.4})$$

omdat dit de reeksontwikkeling van  $1^n = (p+(1-p))^n$  is. Het gemiddelde is

$$\begin{aligned} E[k] &= \sum_{k=0}^n k f(k; n) = \\ &= \sum_{k=1}^n \frac{n(n-1)!}{(k-1)! [(n-1)-(k-1)]!} p^{(k-1)} (1-p)^{[(n-1)-(k-1)]} \\ &= pn \sum_{k'=0}^{n'} \binom{n'}{k'} p^{k'} (1-p)^{n'-k'} \\ &= pn \end{aligned} \quad (\text{C.5})$$

waarbij  $n' = n - 1$  en  $k' = k - 1$ . De variantie vinden we door  $E[k^2] = E[k] + E[k(k-1)]$  uit te rekenen en de bovenstaande truc nog een keer voor  $k-1$  uit te halen. We vinden dan

$$E[k(k-1)] = p^2 n(n-1)$$

en de variantie wordt

$$\text{var}(k) = E[(k-pn)^2] = E[k^2] - p^2 n^2 = p(1-p)n \quad (\text{C.6})$$

Veronderstel dat we in een suspensie van kleine deeltjes het aantal deeltjes per  $\text{cm}^3$  willen bepalen en daarvoor onder een microscoop het aantal deeltjes *tellen* in een monster van  $0,1 \times 0,1 \times 0,1$  mm ( $10^{-6}$   $\text{cm}^3$ ). In een telling treffen we  $k$  deeltjes aan. Aan welke kansverdeling voldoet het aantal deeltjes  $k$  dat we aantreffen?

Laat het gemiddelde aantal deeltjes in het meetvolume  $\mu$  bedragen. Verdeel het meetvolume in een groot aantal  $n$  kleine hokjes, waar hoogstens één deeltje in past. De kans dat een willekeurig hokje bezet is, is  $\mu/n$ .

De kans dat precies  $k$  deeltjes worden aangetroffen is de binomiaalverdeling  $f(k; n)$  met  $p = \mu/n$ . Het aantal deeltjes dat we gemiddeld zullen aantreffen is dus  $pn = \mu$ .

Nu laten we het aantal hokjes  $n$  naar oneindig gaan, maar laten  $\mu$  constant. Dus  $p$  gaat tegelijk naar 0, maar zo dat  $pn = \mu$  constant blijft. De binomiaalcoëfficiënt gaat dan naar  $n^k/k!$ , zodat

$$p(k) \rightarrow \frac{n^k}{k!} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k}$$

De rechterterm gaat naar  $e^{-\mu}$  omdat  $n - k \rightarrow n$  en

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu}, \quad (\text{C.7})$$

zodat

$$f(k) = \frac{\mu^k e^{-\mu}}{k!} \quad (\text{C.8})$$

Dit is de *Poissonverdeling* voor  $k$ , bij gegeven gemiddelde  $\mu$ . Dit is uiteraard een *discrete* verdeling, waarbij  $k$  de gehele waarden  $0, 1, 2, \dots$  kan aannemen. Het gemiddelde  $\mu$  is een parameter van de verdeling en hoeft geen geheel getal te zijn.

Het is gemakkelijk aan te tonen dat de Poissonverdeling genormeerd is en dat zijn gemiddelde gelijk is aan  $\mu$ . Ga dat na en gebruik de reeksontwikkeling

$$e^\mu = \sum_{k=0}^{\infty} \frac{\mu^k}{k!}. \quad (\text{C.9})$$

De variantie van de verdeling is

$$\text{var}(k) = \sigma^2 = E[(k - \mu)^2] = \mu. \quad (\text{C.10})$$

Dit volgt uit  $\sum_{k=0}^{\infty} k^2 \mu^k / k! = \mu^2 + \mu$ , maar is ook de limiet van vgl C.6 voor  $p \rightarrow 0$ .

Bij grote waarden van  $\mu$  benadert de Poissonverdeling een normale verdeling met gemiddelde  $\mu$  en s.d.  $\sqrt{\mu}$ . Bij de afleiding hiervan moeten we erg oppassen dat we tot de juiste orde in de benadering gaan vanwege het wegvallen van termen. We laten zowel  $k$  als  $\mu$  naar  $\infty$  gaan, definiëren

$$x = \frac{k - \mu}{\sqrt{\mu}}; \quad k = \mu + x\sqrt{\mu},$$

en gebruiken de Stirling benadering van  $k!$ :

$$k! = k^k e^{-k} \sqrt{2\pi k} [1 + O(k^{-1})] \quad (\text{C.11})$$

Hiermee wordt

$$\begin{aligned} \ln f(k) &= k - \mu - k \ln(k/\mu) - \frac{1}{2} \ln(2\pi k) + O(k^{-1}) \\ &= x\sqrt{\mu} - (\mu + x\sqrt{\mu} + \frac{1}{2}) \ln\left(1 + \frac{x}{\sqrt{\mu}}\right) - \frac{1}{2} \ln(2\pi\mu) \end{aligned}$$

De laatste term gaat logaritmisches naar  $-\infty$ . We verwachten zo'n term omdat we de waarschijnlijkheid  $f(k)$  van één  $k$  uitrekenen (die naar nul gaat!) en niet de waarschijnlijkheidsdichtheid  $f(x)$ . Ontwikkelen we de logaritme

$$\ln(1+z) = z - \frac{1}{2}z^2 + O(z^3), \quad (\text{C.12})$$

dan vinden we uiteindelijk

$$\lim_{k \rightarrow \infty} \ln f(k) = -\frac{1}{2}x^2 - \frac{1}{2} \ln(2\pi\mu).$$

Tussen  $x$  en  $x + dx$  liggen  $\sqrt{\mu} dx$  gehele getallen, zodat

$$f(x) dx = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \quad (\text{C.13})$$

hetgeen te bewijzen was.

## Bijlage D

# Student's t-verdeling

In deze appendix gaan we de waarschijnlijkheidsverdeling na van een *reeks* normaal verdeelde variabelen  $x_i, i = 1, \dots, n$ , met verwachting  $\mu$  en s.d.  $\sigma$ . Eerst leiden we de verdeling  $\text{Prob}(x_1, \dots, x_n | \sigma)$  af voor bekende  $\sigma$ , en daarna de verdeling  $\text{Prob}(x_1, \dots, x_n)$  voor onbekende  $\sigma$ . We zullen zien dat dit ons de verdeling voor  $\langle x \rangle$  geeft; in het eerste geval een normale verdeling met variantie  $\sigma/n$  en in het tweede geval de Student's t-verdeling voor  $n - 1$  vrijheidsgraden.

Stel we hebben een reeks onafhankelijke steekproeven  $x_1, \dots, x_n$  uit een normale verdeling met verwachting  $\mu$  en variantie  $\sigma^2$ . Dan is

$$\begin{aligned} \text{Prob}(x_1, \dots, x_n | \sigma) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\ &= (\sigma\sqrt{2\pi})^{-n} \exp\left[-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right]. \end{aligned} \quad (\text{D.1})$$

Dit kan herschreven worden als

$$(\sigma\sqrt{2\pi})^{-n} \exp\left[-\frac{(\langle x \rangle - \mu)^2 + \langle (\Delta x)^2 \rangle}{2\sigma^2/n}\right]. \quad (\text{D.2})$$

Het is interessant om te zien dat de volledige waarschijnlijkheid van de serie (vgl D.1) gegeven is door alleen het gemiddelde en de kwadratisch gemiddelde afwijking van het gemiddelde van de serie: deze twee gegevens zijn dus voldoende ('sufficient statistics') om alles over de statistiek van de serie te weten. Maar dat geldt alleen voor normaal verdeelde variabelen!

Voor bekende  $\sigma$  zien we uit formule D.2 dat  $\langle x \rangle$  normaal verdeeld is rond  $\mu$  met variantie  $\sigma^2/n$ .

Voor onbekende  $\sigma$  moeten we integreren over alle mogelijke waarden van  $\sigma$  (we krijgen dan een zg *marginale* verdeling). Het is willekeurig of we hierbij als onbekende variabele  $\sigma$  zelf, de variantie,  $1/\sigma$ , of een andere willekeurige macht van  $\sigma$  nemen; omdat het antwoord niet van zo'n *scaling* mag afhangen, integreren we over  $\log \sigma$ , hetgeen neerkomt op het integreren met  $d\sigma/\sigma$ :

$$\text{Prob}(x_1, \dots, x_n) = \int_0^\infty \text{Prob}(x_1, \dots, x_n | \sigma) \frac{1}{\sigma} d\sigma. \quad (\text{D.3})$$

Deze integraal is evenredig met

$$\int_0^\infty \sigma^{-(n+1)} \exp\left(-\frac{q}{\sigma^2}\right) d\sigma$$

waar

$$q = \frac{1}{2}n[\langle x \rangle - \mu]^2 + \langle (\Delta x)^2 \rangle.$$

Substitueren we  $q/\sigma^2$  door een nieuwe variabele, dan blijkt dat de integraal evenredig wordt met

$$\text{Prob}(x_1, \dots, x_n) \propto \left(1 + \frac{\langle x \rangle - \mu}{\langle (\Delta x)^2 \rangle}\right)^{-n/2} \quad (\text{D.4})$$

Dit is (op een evenredigheidsfactor na, die uit normering wordt verkregen) precies de Student t-verdelingsfunctie  $f(t|\nu)$  voor  $\nu = n - 1$  vrijheidsgraden, met als variabele

$$t = \sqrt{\frac{n(n-1)\langle x \rangle - \mu}{\langle (\Delta x)^2 \rangle}} = \frac{\langle x \rangle - \mu}{\hat{\sigma}/n}, \quad (\text{D.5})$$

waarbij  $\hat{\sigma}^2 = [n/(n-1)] \langle (\Delta x)^2 \rangle$

## Bijlage E

# Schatting van de variantie

*Waarom is de schatting van de variantie groter dan het gemiddelde van de kwadratische afwijkingen van het gemiddelde?*

Hiervoor is het nodig om de verwachting van  $\langle(\Delta x)^2\rangle$  te berekenen. Realiseren we ons eerst dat

$$\begin{aligned}\langle(\Delta x)^2\rangle &= \langle(x - \mu - \langle x - \mu \rangle)^2\rangle \\ &= \sigma^2 - \langle x - \mu \rangle^2\end{aligned}$$

dan zien we dat

$$\begin{aligned}E[\langle(\Delta x)^2\rangle] &= \sigma^2 - E\left[\left(\frac{1}{n}\sum_{i=1}^n(x_i - \mu)\right)^2\right] \\ &= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n E[(x_i - \mu)(x_j - \mu)] \\ &= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^n E[(x_i - \mu)^2] \\ &= \sigma^2\left(1 - \frac{1}{n}\right).\end{aligned}\tag{E.1}$$

De dubbelsom vereenvoudigt hier tot een enkelsom omdat  $x_i$  en  $x_j$  onafhankelijk zijn en alleen de term  $j = i$  in de tweede som overblijft.

Hieruit volgt dat de beste schatting van  $\sigma^2$  gelijk is aan  $n/(n-1)$  maal het gemiddelde van de kwadratische afwijkingen van het gemiddelde.

We zien uit de afleiding dat expliciet gebruik gemaakt is van de aanname dat de statistische afwijkingen van de meetwaarden onafhankelijk van elkaar zijn. Is dat niet het geval dan blijven er in de dubbelsom over  $E[(x_i - \mu)(x_j - \mu)]$  ook termen voor  $j \neq i$  over en wordt de term die van  $\sigma^2$  afgetrokken moet worden, nog groter. Hieronder volgt de formule voor het geval dat er een bekende correlatie tussen *opvolgende* meetpunten is. Van zo'n formule wordt zelden gebruik gemaakt (hij is in feite moeilijk in de literatuur te vinden).<sup>1</sup>

De formule voor de dubbelsom is als volgt:

$$\frac{1}{n^2} \sum_i \sum_j E[(x_i - \mu)(x_j - \mu)] = \sigma^2 \frac{n_c}{n}, \quad (\text{E.2})$$

waar  $n_c$  een soort *correlatielengte* is:

$$n_c = 1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \rho_k, \quad (\text{E.3})$$

$\rho_k$  is de correlatiecoëfficiënt tussen  $x_i$  en  $x_{i+k}$ :

$$E[(x_i - \mu)(x_{i+k} - \mu)] = \rho_k \sigma^2 \quad (\text{E.4})$$

De aanname is gemaakt dat de geordende reeks  $x_i$  een *stationaire* stochastische variabele is (dus met constante variantie en met een correlatiecoëfficiënt die niet van  $i$  afhangt). Voor lange reeksen met korte correlatielengte ( $k \ll n$ ) kan de formule vereenvoudigd worden tot

$$n_c = 1 + 2 \sum_{k=1}^{\infty} \rho_k \quad (\text{E.5})$$

Formule E.2 en E.3 volgen eenvoudig uit het tellen hoe vaak elke term met  $\rho_k$  (vgl. E.4) voorkomt in de dubbelsom.

We krijgen nu in plaats van formule E.1:

$$E[\langle(\Delta x)^2\rangle] = \sigma^2 \left(1 - \frac{n_c}{n}\right). \quad (\text{E.6})$$

---

<sup>1</sup>T.P. Straatsma, H.J.C. Berendsen en A.J. Stam, *Estimation of statistical errors in molecular simulation calculations*, Mol. Phys. **57** (1986) 89.

Het effect van correlatie in de meetpunten op de schatting van  $\sigma^2$  is niet groot en mag ook wel verwaarloosd worden. Maar correlatie heeft wèl een groot effect op de schatting van de variantie in het gemiddelde (zie Appendix F).





## Bijlage F

# Standaarddeviatie van het gemiddelde

*Waarom is de variantie van het gemiddelde  $\langle x \rangle$  van een serie van  $n$  onafhankelijke meetpunten gelijk aan de variantie van  $x$  zelf gedeeld door  $n$ ?*

We vragen naar de volgende grootte:

$$\text{var}(\langle x \rangle) = E[(\langle x \rangle - \mu)^2] = E[\langle x - \mu \rangle^2] \quad (\text{F.1})$$

want  $E[\langle x \rangle] = E[x] = \mu$ . Wanneer we het gemiddelde uitschrijven als een som en het kwadraat als een dubbelsom, krijgen we

$$\text{var}(\langle x \rangle) = \frac{1}{n^2} \sum_i \sum_j E[(x_i - \mu)(x_j - \mu)].$$

Zijn  $x_i$  en  $x_j$  onafhankelijk, dan blijft alleen  $j = i$  over in de tweede som (net als in Appendix E) en wordt

$$\text{var}(\langle x \rangle) = \sigma^2/n. \quad (\text{F.2})$$

*Hoe verandert dit resultaat als de meetpunten afhankelijk zijn?*

Hiervoor moet de bovenstaande dubbelsom uitgewerkt worden. Dit is al gebeurd in Appendix E, vgl E.2, voor het geval van een geordende reeks waarin correlaties alleen afhangen van  $k = j - i$ . We vinden nu:

$$\text{var}(\langle x \rangle) = \sigma^2 \frac{n_c}{n}, \quad (\text{F.3})$$

waar  $n_c$  de correlatielengte is die in Appendix E is gedefiniëerd (formule E.5).

We zien dus dat correlatie de fout in het gemiddelde groter maakt; het effect is alsof we minder meetpunten zouden hebben gehad. Verwaarlozing van de correlatie kan gemakkelijk leiden tot overschatting van de nauwkeurigheid van het gemiddelde. Om een juiste foutenschatting te maken moeten we de correlatielengte  $n_c$  kennen of afleiden uit de meetgegevens. Dit is in het algemeen niet eenvoudig omdat de correlatiecoëfficiënten tussen meetpunten, vooral bij grote afstanden, niet nauwkeurig zijn vast te stellen.

Als alternatieve (en veel eenvoudiger) methode kan ook de volgende *block average* procedure<sup>1</sup> gevolgd worden: groepeer een aantal opvolgende meetpunten die onderling afhankelijk zijn bij elkaar en beschouw het gemiddelde daarvan als één nieuw meetpunt. Deze nieuwe meetpunten zijn dan in veel grotere mate onafhankelijk en kunnen met de normale formules verwerkt worden. Heb je bijv. 1000 meetpunten en verwacht je dat de correlatie zich over enkele tientallen meetpunten uitstrekt, neem dan 10 blokken van 100 punten en beschouw het gemiddelde van elk blok als een onafhankelijk meetpunt. Nog beter is het de bloklengte te variëren en te kijken of de resultaten een betrouwbare limiet hebben. Deze ‘block average’ procedure is niet exact omdat er tussen twee opvolgende blokken altijd nog enige correlatie zal blijven bestaan, maar is wel erg praktisch.

*Hoe groot is de nauwkeurigheid van de geschatte standaarddeviatie?*

Aangezien de variantie van een verdeling geschat wordt uit de som van de kwadratische afwijkingen (gedeeld door  $(n - 1)$ ), voldoet de statistiek van de variantie aan de statistiek van een som van kwadraten. Deze is gegeven door de chi-kwadraatverdeling, waarvan bekend is dat die een gemiddelde  $\nu$  en een variantie  $2\nu$  heeft, dus een *relatieve* s.d. van  $\sqrt{2/\nu}$ , waar  $\nu$  het aantal vrijheidsgraden  $\nu = n - 1$  is. Dus is de relatieve s.d. van de variantie gelijk aan  $\sqrt{2/(n - 1)}$  en de relatieve s.d. van de standaarddeviatie zelf gelijk aan  $1/\sqrt{2(n - 1)}$ . Dit geldt voor onafhankelijke waarnemingen.

---

<sup>1</sup>Ook wel ‘jackknife procedure’ genoemd.

## Bijlage G

# Weegfactoren bij ongelijke varianties

Als we het ‘beste’ gemiddelde willen bepalen van een aantal grootheden  $x_i$ , die gelijke verwachtingen  $\mu$  maar ongelijke standaarddeviaties  $\sigma_i$  hebben, hoe moet de middeling dan plaatsvinden?

Dit moet met gewogen middeling:

$$\langle x \rangle = \frac{1}{w} \sum_{i=1}^n w_i x_i; \quad w = \sum_{i=1}^n w_i. \quad (\text{G.1})$$

De vraag is hoe we de  $w$ 's moeten kiezen: aan welke eis moet de ‘beste’ middeling voldoen? Het criterium dat de schatting ‘zuiver’ moet zijn is hier onbruikbaar, want bij elke keus van de  $w$ 's is de verwachting van het gemiddelde gelijk aan de verwachting van  $x$ . De voor de hand liggende keus is de *minimale variantieschatting*. Dat is de ‘scherpste’ of nauwkeurigste waarde. Laten we dus  $w_i$  zodanig bepalen dat

$$E[(\langle x \rangle - \mu)^2] = E[\langle x - \mu \rangle^2] \text{ minimaal} \quad (\text{G.2})$$

of

$$E[\langle x - \mu \rangle^2] = E \left[ \frac{1}{w^2} \left( \sum_i w_i (x_i - \mu) \right)^2 \right]$$

$$\begin{aligned}
&= \frac{1}{w^2} \sum_{i,j} E[(x_i - \mu)(x_j - \mu)] \\
&= \frac{1}{w^2} \sum_{i,j} w_i w_j \operatorname{cov}(x_i, x_j) = \textit{minimaal} \quad (\text{G.3})
\end{aligned}$$

Veronderstel nu dat  $x_i$  en  $x_j$  onafhankelijk zijn, zodat  $j = i$ , dan zoeken we het minimum van de grootte

$$\sum_i w_i^2 \sigma_i^2$$

onder de conditie dat  $\sum_i w_i$  constant blijft. De standaardmethode om zo'n *optimalisatie met nevenvoorwaarde* op te lossen is de methode van de onbepaalde multiplicator(en) van Lagrange. Daarbij wordt de nevenvoorwaarde ( $\sum_i w_i$  is constant) met een onbepaalde constante  $\lambda$  vermenigvuldigd en toegevoegd aan de te minimaliseren functie. De partiële afgeleiden van de totale functie naar elk van de variabelen wordt vervolgens nul gesteld. De oplossing bevat dan nog deze onbepaalde multiplicator, maar die volgt uit de nevenvoorwaarde. Alsvolgt:

$$\frac{\partial}{\partial w_i} \left( \sum_j w_j^2 \sigma_j^2 + \lambda \sum_j w_j \right) = 2w_i \sigma_i^2 + \lambda = 0$$

Dus

$$w_i \propto \frac{1}{\sigma_i^2}. \quad (\text{G.4})$$

Dus moet het gewicht van een meetpunt omgekeerd evenredig zijn met zijn variantie. Dit geldt als de meetwaarden onafhankelijke fouten hebben.

*Hoe groot is nu de variantie in  $\langle x \rangle$ ?*

Dit bepalen we door de verwachting van  $(\langle x \rangle - \mu)^2$  te berekenen:

$$\sigma_{\langle x \rangle}^2 = E[(\langle x \rangle - \mu)^2] = \frac{1}{w^2} \sum_i w_i^2 \sigma_i^2,$$

waarbij gebruik gemaakt is van de onafhankelijkheid van  $x_i$  en  $x_j$ . Voor  $w_i$  mogen we kiezen  $w_i = 1/\sigma_i^2$ , zodat

$$\sigma_{\langle x \rangle}^2 = \frac{1}{w^2} \sum_i \frac{1}{\sigma_i^2} = \left( \sum_i \frac{1}{\sigma_i^2} \right)^{-1}. \quad (\text{G.5})$$

## Bijlage H

# Kleinste- kwadratenaanpassing

*Hoe vinden we de oplossing voor  $a$  en  $b$  in  $y \approx ax + b$ ?*

De waarden voor  $a$  en  $b$  in de functie  $f(x) = ax + b$ , zodanig dat

$$S = \sum_{i=1}^n w_i (y_i - f_i)^2 = \sum_{i=1}^n w_i (y_i - ax_i - b)^2$$

minimaal is, worden eenvoudig gevonden uit het nul stellen van de afgeleiden van  $S$  naar  $a$  en  $b$ :

$$\begin{aligned} \frac{\partial S}{\partial a} &= -2 \sum_{i=1}^n w_i x_i (y_i - ax_i - b) = 0 \\ \frac{\partial S}{\partial b} &= -2 \sum_{i=1}^n w_i (y_i - ax_i - b) = 0 \end{aligned}$$

Uit de tweede vergelijking volgt dat  $b = \langle y \rangle - a \langle x \rangle$  en wanneer dit in de eerste vergelijking ingevuld wordt vinden we de oplossing voor  $a$ . De gemiddelden zijn met  $w_i$  gewogen

*De algemene lineaire regressie*

In het volgende gebruiken we matrixnotatie. Een (vetgedrukte) kleine letter stelt een kolomvector voor; een (vetgedrukte) hoofdletter een matrix.  $\mathbf{A}^T$  is de getransponeerde van  $\mathbf{A}$ ;  $\text{Tr}(\mathbf{A})$  is het spoor ('trace', som

van de diagonaalelementen) van  $\mathbf{A}$ . Bedenk dat  $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$  en dat het spoor van een matrixproduct invariant is voor cyclische verwisseling van de termen in het product.

In het algemeen kan een stelsel van lineaire relaties in  $m$  parameters  $\theta_k, k = 1, \dots, m$  worden geschreven als

$$f_i(\theta_1, \theta_2, \dots, \theta_m) = \sum_{k=1}^m a_{ik} \theta_k, \text{ of } \mathbf{f}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta}. \quad (\text{H.1})$$

We beschouwen het algemene geval waarin de fluctuaties van  $y$  gecorreleerd mogen zijn. Veronderstel dat de ‘echte’ meetwaarden  $y_i$  gegeven zijn door

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta}_m + \boldsymbol{\epsilon}, \quad (\text{H.2})$$

waar  $\boldsymbol{\theta}_m$  de ‘echte’ *modelwaarden* van de parameters zijn en  $\boldsymbol{\epsilon}$  de toegevoegde stochast of ‘ruis’, met de volgende eigenschappen:

$$E[\boldsymbol{\epsilon}] = \mathbf{0} \quad (\text{H.3})$$

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \boldsymbol{\Sigma} \quad (\text{H.4})$$

$\boldsymbol{\Sigma}$  is de *covariantiematrix* van fouten  $\boldsymbol{\epsilon}$  in de meetwaarden  $\mathbf{y}$ .

De som van kwadraten kan nu geschreven worden als

$$\chi^2 = (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^T \mathbf{W}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}), \quad (\text{H.5})$$

waar de matrix van gewichtsfactoren  $\mathbf{W}$  de inverse van de covariantiematrix van de fouten in de meetwaarden is:

$$\mathbf{W} = \boldsymbol{\Sigma}^{-1}. \quad (\text{H.6})$$

Als er geen correlaties tussen de fouten in de meetwaarden zijn, dan zijn zowel  $\boldsymbol{\Sigma}$  als  $\mathbf{W}$  diagonaal met  $\sigma_i^2$ , resp.  $\sigma_i^{-2}$  als diagonaalelementen.

Voeren we nu de differentiaties uit dan vinden we

$$\frac{\partial \chi^2}{\partial \boldsymbol{\theta}} = -2\mathbf{A}^T \mathbf{W}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) = 0 \quad (\text{H.7})$$

zodat de kleinste-kwadratenoplossing voor  $\boldsymbol{\theta}$ , die we aanduiden met  $\hat{\boldsymbol{\theta}}$ , de oplossing is van het stelsel vergelijkingen

$$\mathbf{B}\boldsymbol{\theta} = \mathbf{A}^T \mathbf{W}\mathbf{y}, \quad (\text{H.8})$$

waar

$$\mathbf{B} = \mathbf{A}^T \mathbf{W} \mathbf{A}. \quad (\text{H.9})$$

Dus is

$$\hat{\boldsymbol{\theta}} = \mathbf{B}^{-1} \mathbf{A}^T \mathbf{W} \mathbf{y}. \quad (\text{H.10})$$

Hiermee kunnen dus algemene lineaire regressies worden uitgevoerd. Merk op dat de absolute waarden van de gewichtsfactoren niet belangrijk zijn voor de kleinste-kwadratenoplossing: als alle waarden van  $\mathbf{W}$  met een constante vermenigvuldigd worden, verandert dat de oplossing voor  $\hat{\boldsymbol{\theta}}$  niet.

De gevonden kleinste-kwadratenwaarde van  $\hat{\boldsymbol{\theta}}$  is een *zuivere schatting* van  $\boldsymbol{\theta}$ , omdat

$$E[\hat{\boldsymbol{\theta}}] = \mathbf{B}^{-1} \mathbf{A}^T \mathbf{W} E[\mathbf{y}] = \boldsymbol{\theta}_m \quad (\text{H.11})$$

omdat  $E[\mathbf{y}] = \mathbf{A} \boldsymbol{\theta}_m$ .

*Chi-kwadraat als functie van de parameters*

De uitdrukking H.5 voor  $\chi^2(\boldsymbol{\theta})$  kunnen we ook schrijven als een kwadratische functie in de parameters. Definiëren we voor de afwijking van de geschatte parameters

$$\boldsymbol{\Delta} \boldsymbol{\theta} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}, \quad (\text{H.12})$$

en voor het minimum van  $\chi^2$

$$\chi_0^2 = (\mathbf{y} - \mathbf{A} \hat{\boldsymbol{\theta}})^T \mathbf{W} (\mathbf{y} - \mathbf{A} \hat{\boldsymbol{\theta}}) \quad (\text{H.13})$$

en vullen we H.10 en H.12 in H.5 in, dan vinden we met gebruikmaking van H.7 dat

$$\chi^2(\boldsymbol{\theta}) = \chi_0^2 + \boldsymbol{\Delta} \boldsymbol{\theta}^T \mathbf{B} \boldsymbol{\Delta} \boldsymbol{\theta}. \quad (\text{H.14})$$

$\chi^2$  is dus een parabolische functie in  $\boldsymbol{\Delta} \boldsymbol{\theta}$  en de matrix van tweede afgeleiden is gegeven door  $2 \mathbf{B}$ .

*Covarianties van de parameters*

De covariantiematrix  $\text{cov}(\hat{\theta}_k, \hat{\theta}_l)$  van de geschatte parameters is gegeven door  $E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_m)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_m)^T]$ . Voor het gemak definiëren we

$$\mathbf{C} = \mathbf{B}^{-1} \quad (\text{H.15})$$

Omdat

$$\boldsymbol{\theta}_m = \mathbf{C} \mathbf{A}^T \mathbf{W} \mathbf{A} \boldsymbol{\theta}_m$$

is (met vgl H.10)

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_m = \mathbf{C} \mathbf{A}^T \mathbf{W} (\mathbf{y} - \mathbf{A} \boldsymbol{\theta}_m) = \mathbf{C} \mathbf{A}^T \mathbf{W} \boldsymbol{\epsilon}. \quad (\text{H.16})$$



We vinden nu voor de covariantiematrix

$$\begin{aligned} E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_m)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_m)^T] &= E[\mathbf{CA}^T \mathbf{W} \boldsymbol{\epsilon} (\mathbf{CA}^T \mathbf{W} \boldsymbol{\epsilon})^T] \\ &= \mathbf{CA}^T \mathbf{W} E[\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T] \mathbf{W} \mathbf{A} \mathbf{C} \\ &= \mathbf{C} \end{aligned} \quad (\text{H.17})$$

omdat  $E[\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T] = \boldsymbol{\Sigma}$ . We zien dus dat de matrix  $\mathbf{C}$  de covarianties van de geschatte parameters geeft.

De verwachte waarde van  $\chi_0^2$

De verwachte waarde van  $\chi_0^2$  kunnen we, door  $\mathbf{y}$  uit vgl H.2 in te vullen in H.13, schrijven als

$$E[\chi_0^2] = E[\{\mathbf{A}(\boldsymbol{\theta}_m - \hat{\boldsymbol{\theta}}) + \boldsymbol{\epsilon}\}^T \mathbf{W} \{\mathbf{A}(\boldsymbol{\theta}_m - \hat{\boldsymbol{\theta}}) + \boldsymbol{\epsilon}\}]. \quad (\text{H.18})$$

Vullen we hierin vgl H.16 in, dan krijgen we

$$E[\chi_0^2] = E[\boldsymbol{\epsilon}^T (\mathbf{I} - \mathbf{A} \mathbf{C} \mathbf{A}^T \mathbf{W})^T \mathbf{W} (\mathbf{I} - \mathbf{A} \mathbf{C} \mathbf{A}^T \mathbf{W}) \boldsymbol{\epsilon}], \quad (\text{H.19})$$

waar  $\mathbf{I}$  de  $n \times n$  eenheidsmatrix is. Deze expressie heeft de vorm

$$E[\boldsymbol{\epsilon}^T \mathbf{M} \boldsymbol{\epsilon}] = \sum_{i,j} E[\epsilon_i \epsilon_j M_{ij}] = \text{Tr}(\boldsymbol{\Sigma} \mathbf{M}) \quad (\text{H.20})$$

en uitwerken in termen geeft

$$E[\chi_0^2] = \text{Tr}(\boldsymbol{\Sigma} \mathbf{W}) - \text{Tr}(\boldsymbol{\Sigma} \mathbf{W} \mathbf{A} \mathbf{C} \mathbf{A}^T \mathbf{W}) \quad (\text{H.21})$$

De eerste term is  $\text{Tr}(\mathbf{I}) = n$  en de tweede term is gelijk aan  $\text{Tr}(\mathbf{A} \mathbf{C} \mathbf{A}^T \mathbf{W}) = \text{Tr}(\mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{C}) = m$  omdat dit het spoor van een  $m \times m$  eenheidsmatrix is. Dus is de verwachte waarde van het minimum van  $\chi^2$  gelijk aan het aantal meetpunten verminderd met het aantal parameters (of: het aantal vrijheidsgraden  $\nu$ ):

$$E[\chi_0^2] = \nu = n - m. \quad (\text{H.22})$$

*Waarom is de s.d. van een parameter gegeven door de projectie van de ellipsoïde  $\Delta\chi^2 = 1$ ?*

De conditie  $\Delta\chi^2 = 1$  beschrijft een oppervlak (een ellipsoïde) in de  $m$ -dimensionale parameter ruimte. In figuur 5.2 is met raaklijnen aan de ellips  $\Delta\chi^2 = 1$  aangegeven dat de *projectie* van deze figuur op één van de assen (bijv.  $\theta_1$ ) de grenzen  $\hat{\theta}_1 \pm \sigma_{\theta_1}$  aftekent. De raaklijn gaat door

een punt van de ellips waar  $\chi^2$  minimaal is t.o.v. alle *andere* parameters  $\theta_2, \dots, \theta_m$ , dus waar de gradient van  $\chi^2$  in de richting van  $\theta_1$  staat:

$$\mathbf{grad} \chi^2 = (a, 0, \dots, 0)^T,$$

waar  $a$  een constante is die volgt uit  $\Delta\chi^2 = \mathbf{\Delta\theta}^T \mathbf{B} \mathbf{\Delta\theta} = 1$ : omdat<sup>1</sup>  $\mathbf{grad} \chi^2 = 2\mathbf{B} \mathbf{\Delta\theta}$ , is

$$\mathbf{\Delta\theta}^T \frac{1}{2}(a, 0, \dots, 0)^T = \frac{1}{2}a\Delta\theta_1 = 1.$$

Dus is

$$\mathbf{B} \mathbf{\Delta\theta} = \frac{1}{2}(2/\Delta\theta_1, 0, \dots, 0)^T$$

en is

$$\mathbf{\Delta\theta} = \mathbf{C}(1/\Delta\theta_1, 0, \dots, 0)^T \text{ of } \Delta\theta_1 = \pm\sqrt{C_{11}} = \pm\sigma_{\theta_1}. \quad (\text{H.23})$$

hetgeen te bewijzen was. Dit bewijs kan gevonden worden in ref. [8].

*De niet-lineaire kleinste-kwadratenaanpassing*

Als de functies  $f_i(\theta_1, \dots, \theta_m)$  niet lineair in alle parameters zijn, maar  $\chi^2 = (\mathbf{y} - \mathbf{f})^T \mathbf{W}(\mathbf{y} - \mathbf{f})$  heeft wel een minimum  $\chi_0^2 = \chi^2(\hat{\boldsymbol{\theta}})$ , dan kan  $\chi^2(\boldsymbol{\theta})$  rond dat minimum worden ontwikkeld in een Taylorreeks, waarbij de lineaire term nul is:

$$\chi^2(\boldsymbol{\theta}) = \chi_0^2 + \mathbf{\Delta\theta}^T \mathbf{B} \mathbf{\Delta\theta} + \dots, \quad (\text{H.24})$$

net als in vgl H.14 in het lineaire geval. Met de herdefinitie van de matrix  $\mathbf{A}$ :

$$A_{ik} = \left( \frac{\partial f_i}{\partial \theta_k} \right)_{\hat{\boldsymbol{\theta}}} \quad (\text{H.25})$$

blijven alle formules voor de parameters en hun varianties bij benadering geldig. De inverse van  $\mathbf{B} = \mathbf{A}^T \mathbf{W} \mathbf{A}$  is ook hier (bij benadering nu) de covariantiematrix van de parameters. Zie [8] voor een discussie hiervan. Voor onafhankelijke meetpunten geldt vgl 5.23 uit par. 3.4:

$$B_{kl} = \sum_{i=1}^n \sigma_i^{-2} \frac{\partial f_i}{\partial \theta_k} \frac{\partial f_i}{\partial \theta_l}. \quad (\text{H.26})$$

---

<sup>1</sup>De gradient van een kwadratische vorm  $\frac{1}{2}\mathbf{x}^T \mathbf{G} \mathbf{x}$ , met  $\mathbf{G}$  symmetrisch, is gelijk aan  $\mathbf{G} \mathbf{x}$ .



# Antwoorden van de opgaven

- 1.1**  $l = 31,3 \pm 0,2$  m (tenzij de precisie werkelijk  $20 \pm 1$  cm bedraagt, dan:  $l = 31,30 \pm 0,20$  m);  $c = 15,3 \pm 0,1$  mM;  $k = 252$  S/m;  $k/M^{-1}s^{-1} = (35,7 \pm 0,7) \times 10^2; = 2,00 \pm 0,03$ .
- 1.2** *a.*  $3,00 \pm 0,06$  (rel. fout 2%); *b.*  $6,0 \pm 0,3$  (rel. fout  $\sqrt{3^2 + 4^2}\%$ ); *c.*  $3,000 \pm 0,001$ . Bedenk dat  $\log_{10}(1 \pm \delta) = \pm 0,434 \ln(1 + \delta) \approx \pm 0,434\delta = 0,00087$ . Bij ingewikkelde functies is het soms handiger om beide grenzen direct te berekenen:  $\log_{10} 998 = 2,99913$  en  $\log_{10} 1002 = 3,00087$ ; *d.*  $2,71 \pm 0,06$  (rel. fout  $\sqrt{1.5^2 + 1^2}\%$ ).
- 1.3**  $k = \ln 2/\text{halfwaardetijd}$ . De relatieve fout in  $k$  is gelijk aan de relatieve fout in de halfwaardetijd. De fout in  $\ln k$  is gelijk aan de relatieve fout in  $k$ :  $\sigma(\ln k) = \sigma(k)/k$ . De volgende waarden worden verkregen:

$\frac{1000}{T/K}$	$k/s^{-1}$	$\ln k/s^{-1}$
1,2772	$(0,347 \pm 0,017) \times 10^{-3}$	$-7,97 \pm 0,05$
1,2300	$(1,155 \pm 0,077) \times 10^{-3}$	$-6,76 \pm 0,07$
1,1862	$(2,89 \pm 0,24) \times 10^{-3}$	$-5,85 \pm 0,08$
1,1455	$(7,70 \pm 0,86) \times 10^{-3}$	$-4,87 \pm 0,11$

- 1.4**  $9,80 \pm 0,3$  (Relatieve fout is  $\sqrt{0,2^2 + (2 \times 0,1)^2} = 0,28\%$ )
- 1.5** Aangezien  $\Delta G = RT \ln(kh/k_B T)$ , is de afgeleide naar  $T$  gelijk aan  $(\Delta G/T) + R$ . Dat is  $(30\,000/300) + 8,3 = 108,3$ . Dat betekent dat een afwijking in  $T$  van  $\pm 5$  een afwijking in  $\Delta G$  geeft van  $108,3 \times 5 = 540$  J/mol.

- 2.1** Het antwoord is 3, en dus wordt de kurtosis en niet de excess bedoeld.
- 2.2**  $f(0) = 0,59874$ ;  $f(1) = 0,31512$ ;  $f(2) = 0,074635$ ;  $f(3) = 0,010475$ ;  $f(4) = 0,000965$ .
- 2.3** Gezocht wordt  $1 - f(0) = 1 - 0,99^{20} = 0,182$
- 2.4** Aannemende dat we hier met een Poissonproces te maken hebben, is de standaardfout gelijk aan de wortel uit het waargenomen aantal, dus de lichtmeting geeft  $900 \pm 30$  pulsen en de donkermeting geeft  $100 \pm 10$  pulsen. De lichtintensiteit is dus evenredig met  $(900 - 100) \pm \sqrt{30^2 + 10^2} = 800 \pm 32$ . De relatieve fout is dus 4%. Wordt de meting 100 keer herhaald (of wordt 100 keer langer gemeten) dan worden de gemeten aantallen  $100\times$  zo groot, maar de (absolute) fouten slechts  $10\times$  zo groot. De relatieve fout wordt dus  $10\times$  zo klein (0,4%).
- 2.5** De uniforme verdeling  $f(x) = 1, 0 \leq x < 1$ , heeft gemiddelde 0,5 en variantie  $\sigma^2 = \int_0^1 (x - 0,5)^2 dx = 1/12$ ; 12 keer optellen geeft 12 keer grotere variantie. MATHEMATICA voorbeeld:  
`gauss := Sum[Random[], {i, 12}] - 6.0`  
`r = Table[gauss, {i, 100}]`
- 2.6**  $F(0,1) - F(-0,1) = 2 \times (0,5 - 0,4602) = 0,0796$ . Merk op dat dit vrijwel gelijk is aan  $f(0) \times 0,2 = 0,0798$ .
- 2.7**  $f(6) = 6,076 \times 10^{-9}$ ;  $F(-6) = 1,0125 \times 10^{-9}(11/38 + \dots) = 9,86 \times 10^{-10}$ , waarschijnlijk nauwkeurig op twee decimalen.
- 3.1** Ja (rechte lijn);  $\mu = 5,20$ ;  $\sigma = 1,55$  Afleesnauwkeurigheid van deze getallen ca 0,05.
- 3.2** Werk het kwadraat in  $\frac{1}{n} \sum (x_i - \langle x \rangle)^2$  uit.
- 3.3** Pas de formule toe op  $y = x - c$ .
- 3.4**  $\langle x \rangle = 5,228$ ;  $\langle (\Delta x)^2 \rangle = 2,255$ ;  $\hat{\sigma}^2 = 2,349$ ;  $\hat{\sigma} = 1,533$ ;  $\hat{\sigma}(\langle x \rangle) = 0,307$ ; nauwkeurigheid hierin 0,044. Gemiddelde:  $5,23 \pm 0,31$ .  
 Student's t-verdeling: 50% acceptatie tussen 5,02 en 5,44; 90% acceptatie tussen 4,70 en 5,76.  
 Voor MATHEMATICA kan `Apply` gebruikt worden om een bewerking over een lijst uit te voeren, bijv.:  
`average = Apply[Sum, xlist]/Length[xlist]`

**3.5** Om numerieke fouten te vermijden is het beter alleen met de laatste cijfers van het getal te werken. Nemen we voor  $x_i(\sigma_i)$  het 5e, 6e en 7e cijfer achter de komma, dan hebben we de volgende getallen: 367(036); A: 348(075); B: 372(030); C: 500(500).

Voor de gewichten nemen we  $w_i = 1/\sigma_i^2$ . Toepassing van de formules levert dan 368(022) op. Dus:  $N_{Av} = 6,0221368(22)$ .

**4.1** Lijn gaat door punten (9, 100) en (188,1) (precisie ca 1%). Geeft  $k = \ln 100/(188 - 9) = 0,0257$  en  $c_0 = 126$ .

**4.2** (In te veel decimalen:) Lineweaver-Burk:  $K_m = 1/0,0094 = 106,383$ ;  $v_{max} = K_m(0,04+0,0094)/0,35 = 15,015$ ; Eadie-Hofstee:  $K_m = (15 - 2)/(0,120 - 0,007) = 115,04$ ;  $v_{max} = 0,120K_m + 2 = 15,805$ ; Hanes:  $v_{max} = 500/(39 - 7.5) = 15,873$ ;  $K_m = 7,5v_{max} = 119,05$ .

**4.3** Rechte lijn gaat ongeveer door (1,1, -3,66) en (1,3, -8,55) met verticale precisie ca 0,1. Of:  $E/R = (8,55 - 3,66)/0,2 \times 10^{-3} \pm 2\% = 203 \pm 4$  kJ/mol. De waarde van  $A$  kan door extrapolatie worden verkregen:  $\ln A = -3,66 + (1,2/0,2)(8,55 - 3,66) = 23,24$  De fout in  $\ln A$  (zelfde als relatieve fout in  $A$ ) wordt geheel bepaald door de fout in  $E$  en bedraagt  $(1/RT) \times$  fout in  $E = 0,6$ , dus  $\ln A = 23,2 \pm 0,6$ .

**4.4** De ijklijn gaat door de punten (60, 1,24),(70, 1,40). Interpolatie geeft voor O.D =  $1,38 \pm 0,01$ ;  $c = 60 + 10 \times (1,38 - 1,24)/(1,40 - 1,24) = 68,75 \pm (0,01 \times 10/0,16) = 68,8 \pm 0,6$  mM.

**5.1** Denk eraan dat  $\sigma_y = \sigma_v/v^2$ . Toepassing van **linreg** geeft  $a = 7,23 \pm 0,31$ ;  $b = 0,0636 \pm 0,0017$ ;  $\rho_{ab} = -0,816$ . Hieruit volgt  $v_{max} = 15,72 \pm 0,41$  en  $K_m = a/b = 114 \pm 8$ . De fout in  $K_m$  volgt uit:

$$\frac{\delta K_m}{K_m} = \frac{\delta a}{a} - \frac{\delta b}{b}$$

of

$$\left(\frac{\sigma_{K_m}}{K_m}\right)^2 = \left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_b}{b}\right)^2 - 2\rho_{ab} \left(\frac{\sigma_a}{a}\right) \left(\frac{\sigma_b}{b}\right).$$

De covariantie tussen  $K_m$  en  $v_{max}$  kan ook berekend worden door uitwerking van

$$\langle \delta K_m \delta v_{max} \rangle = \left\langle \delta \left(\frac{a}{b}\right) \delta \left(\frac{1}{b}\right) \right\rangle.$$

De waarde daarvan is 2,95 en de correlatiecoëfficiënt tussen  $K_m$  en  $v_{max}$  wordt 0,95.

- 5.2** Gebruik `Fit` of `linreg` met gewichten gelijk aan 1. Dit levert (in te veel decimalen):  $a = -0,25857$ ;  $b = 32,7429$ ;  $\sigma_a = 0,0131$ ;  $\sigma_b = 3,95$ ;  $\rho_{ab} = -0,997$ . De entropie ( $S = -a$ ) is dus  $259 \pm 13$  J/mol. Extrapolatie naar  $350 \text{ K} = f(350)$  levert  $f = 19,814$  op. Reken je de fout hierin uit als  $\sqrt{\sigma_a^2 x^2 + \sigma_b^2}$  dan krijg je een waarde van 6. Dit antwoord is onjuist door het verwaarlozen van de correlatie. Ga nu uit van  $T - 300$  voor de  $x$ -waarden. De waarden van  $a, b, \sigma_a$  en de voorspelling voor  $350 \text{ K} = f(50)$  blijven hetzelfde, maar  $b$  wordt  $32,743$ ,  $\sigma_b$  wordt  $0,2626$  en  $\rho_{ab} = 0$ . Nu is de fout in de geëxtrapolerde waarde slechts  $0,71$ ! Het verschil komt door de zeer sterke correlatie tussen  $a$  en  $b$ . De foutberekening moet rekening houden met de covariantie door formule 1.4 toe te passen (zie ook Bijlage A):

$$\text{var}(ax + b) = x^2 \text{var}(a) + \text{var}(b) + 2x \text{cov}(a, b),$$

hetgeen de correcte waarde  $0,71$  oplevert. Het juiste antwoord is dus  $\Delta G(350\text{K}) = (19,8 \pm 0,7) \text{ kJ/mol}$ , maar dit wordt eenvoudiger verkregen als de  $T$ -waarden t.o.v. het gemiddelde van  $T$  wordt genomen in de kleinste kwadratenaanpassing.

- 5.3**  $a = 71,464$ ;  $b = 19,140$ ;  $p = 0,098173$ ;  $q = 0,018307$ . Merk op dat deze waarden nogal verschillen van de grafische schatting. Het aanpassen van meerdere e-machten is moeilijk: de parameters zijn sterk onderling afhankelijk en soms kan er geen minimum worden gevonden.
- 5.4**  $\nu = 6 - 2 = 4$ ;  $\chi^2 = 4,3$  ligt dichtbij de 50% waarde en is zeer acceptabel. Met `MATHEMATICA` vind je dat de cumulatieve verdelingsfunctie de waarde  $0,630$  heeft.

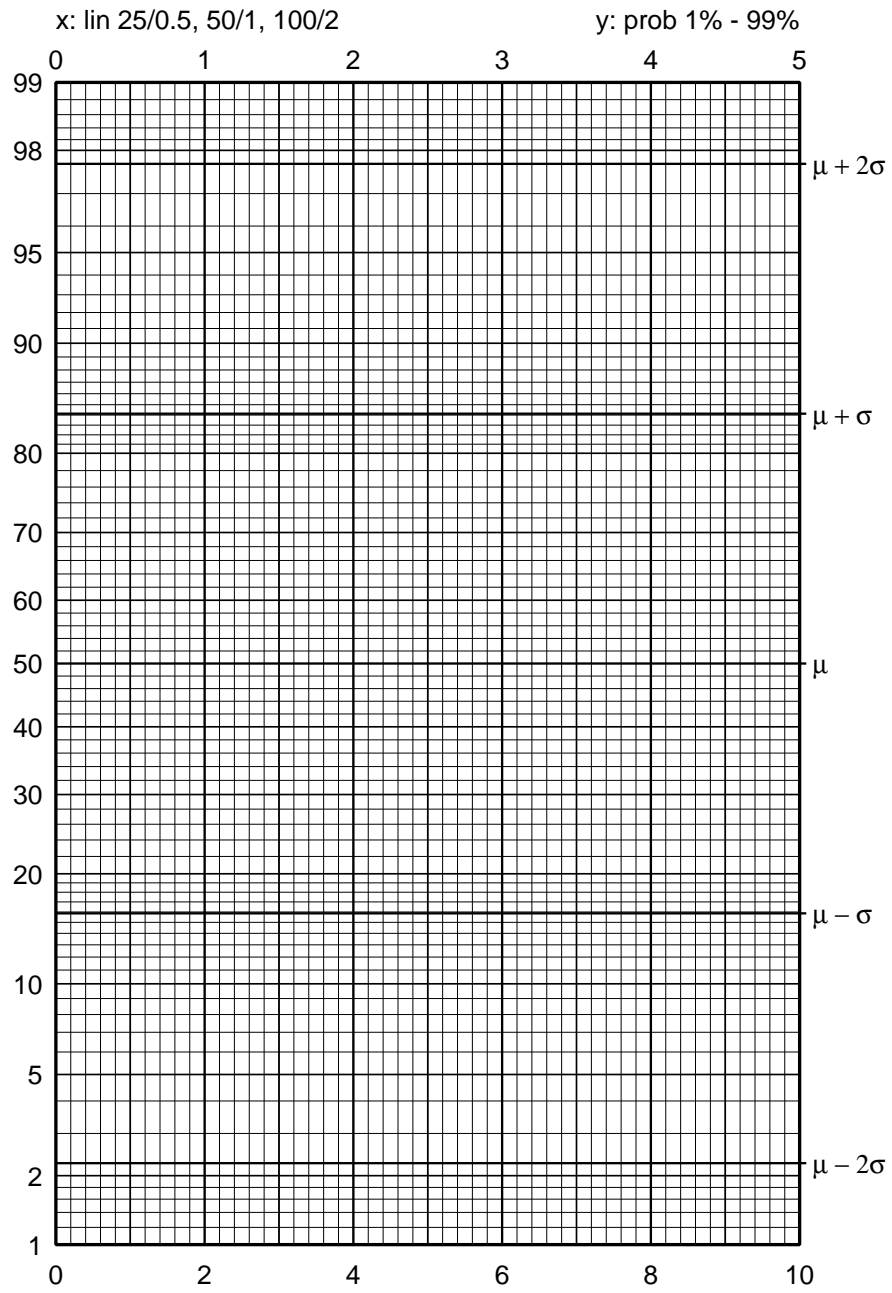
# Waarschijnlijkheidspapier

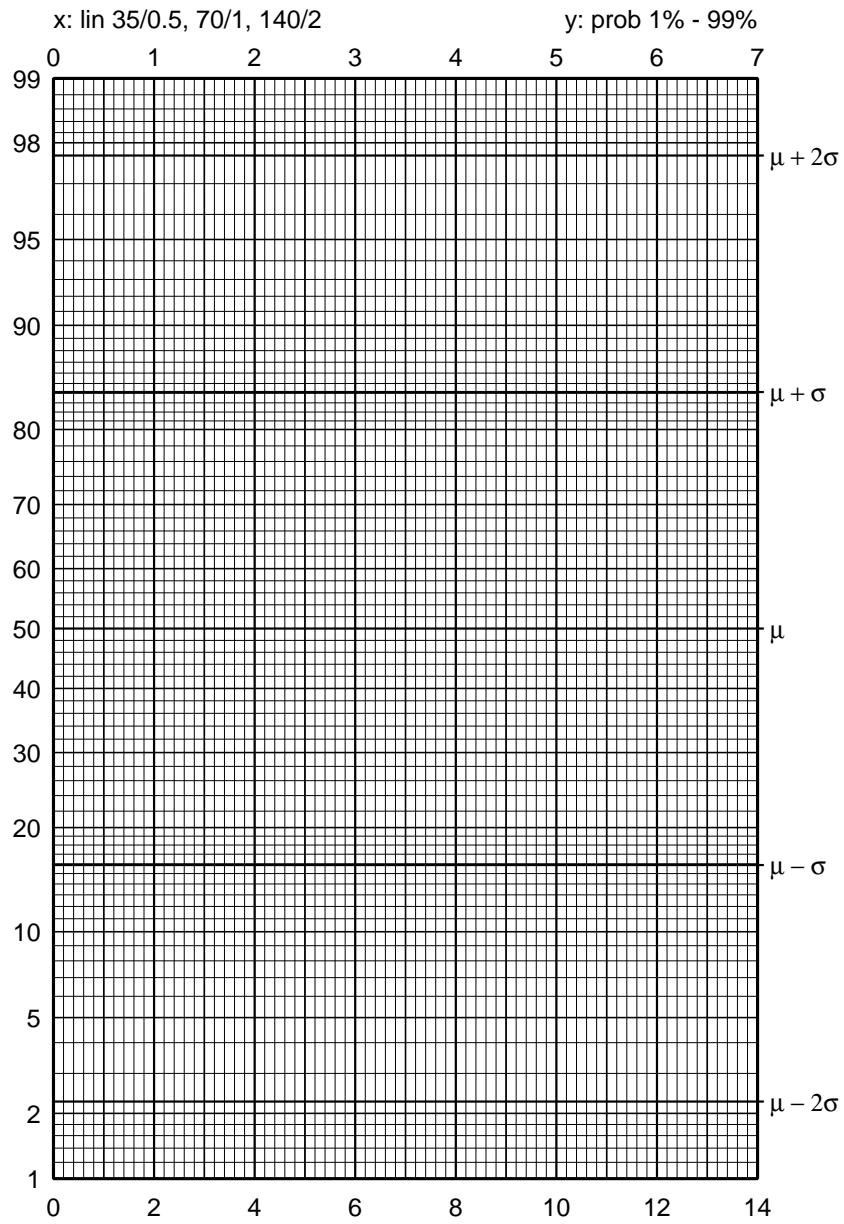
Er zijn twee soorten waarschijnlijkheidspapier bijgevoegd, met verschillende horizontale lineaire indelingen, in resp. 50 en 70 eenheden. De vertikale schaal loopt van 1 % tot 99 %. Copieer deze pagina's voor gebruik.

Men zet het rangnummer van een serie geordende meetwaarden op de vertikale schaal uit (in procenten van het totaal aantal meetwaarden  $n$ ), tegen de meetwaarde zelf op de horizontale schaal. Doe dit 'trapsgewijs': trek een vertikale lijn van  $100(k-1)/n$  % tot  $100k/n$  % bij de waarde  $x_k$  en een horizontale lijn van  $x_k$  tot  $x_{k+1}$  bij het percentage  $100k/n$ . De eerste vertikale lijn komt dus van  $-\infty$  en de laatste gaat naar  $+\infty$ . Zie figuur 3.3 voor een voorbeeld.

Als de waarden een normale verdeling volgen, zal de trapsgewijze figuur gemiddeld op een rechte lijn liggen. Die lijn snijdt de 50% lijn bij de verwachte waarde, terwijl de snijpunten met de dik getrokken lijnen een interval ter breedte van  $2\sigma$  insluiten. Is de beste lijn door de punten niet recht, dan kunnen betrouwbaarheidsintervallen direct van die lijn afgelezen worden.









# Zakboekblaadjes

De volgende zakboekblaadjes zijn bijgevoegd in alfabetische rangschikking:

• Chikwadraatverdeling	115
• Eenheden	117
• Fysische constanten	123
• Kleinste kwadratenaanpassing	125
• Mathematica	127
• Normale verdeling	133
• Student t-verdeling	135

Wanneer de een blad tot 71% verkleind wordt gecopiëerd (van A3 naar A4 formaat) past het blad in het standaardformaat van de Junior Succes agenda.



CHI-KWADRAATVERDELING

1

**KANSVERDELING SOM VAN KWADRATEN**

Stel  $x_1, x_2, \dots, x_\nu$  zijn onafhankelijke, normaal gedistribueerde variabelen met  $E\{x_i\} = 0$  en  $E\{x_i^2\} = 1$ ;  $\nu =$  aantal vrijheidsgraden;  $\chi^2 = \sum_{i=1}^{\nu} x_i^2$ . De waarschijnlijkheidsdichtheid van  $\chi^2$  is:

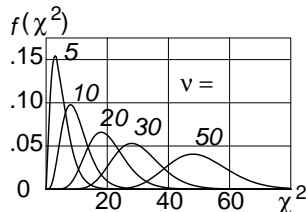
$$f(\chi^2|\nu) d\chi^2 = [2^{\nu/2} \Gamma(\frac{\nu}{2})]^{-1} (\chi^2)^{\nu/2-1} \exp[-\chi^2/2] d\chi^2.$$

**Momenten van  $f(\chi^2|\nu)$ :**

gemiddelde  $\mu = E\{\chi^2\} = \nu$   
 variantie  $\sigma^2 = E\{(\chi^2 - \mu)^2\} = 2\nu$   
 'skewness'  $\gamma_1 = E\{(\chi^2 - \mu)^3/\sigma^3\} = 2\sqrt{2/\nu}$   
 excess  $\gamma_2 = E\{(\chi^2 - \mu)^4/\sigma^4 - 3\} = 12/\nu$

**Bijzondere gevallen**

$\nu$	$f(\chi^2 \nu)$
1	$(2\pi)^{-1/2} \chi^{-1} \exp[-\chi^2/2]$
2	$\frac{1}{2} \exp[-\chi^2/2]$
3	$(2\pi)^{-1/2} \chi \exp[-\chi^2/2]$
$\infty$	$(4\pi\nu)^{-1/2} \exp[-(\chi^2 - \nu)^2/(4\nu)]$ normaal met var = $2\nu$



**Relatie tot Poisson verdeling ( $\nu$  even)**

$$1 - F(\chi^2|\nu) = \sum_{j=0}^{c-1} e^{-m} m^j / j!,$$

$$c = \frac{1}{2}\nu \quad m = \frac{1}{2}\chi^2.$$

**CUMULATIEVE  $\chi^2$ -VERDELING**

$F(\chi^2|\nu) =$  kans dat som van kwadraten  $< \chi^2$ :

$$F(\chi^2|\nu) = \int_0^{\chi^2} f(S|\nu) dS \quad \text{Zie tabel p. 2}$$

Kans dat  $\chi^2$  overschreden wordt is  $1 - F(\chi^2)$

## CHI-KWADRAATVERDELING

2

Waarden van  $\chi^2$  bij 1%, 10%, 50%, 90% en 99%

$F =$ $\nu$	0,01	0,10	0,50	0,90	0,99
1	0,000	0,016	0,455	2,706	6,635
2	0,020	0,211	1,386	4,605	9,210
3	0,115	0,584	2,366	6,251	11,35
4	0,297	1,064	3,357	7,779	13,28
5	0,554	1,610	4,351	9,236	15,09
6	0,872	2,204	5,348	10,65	16,81
7	1,239	2,833	6,346	12,02	18,48
8	1,646	3,490	7,344	13,36	20,09
9	2,088	4,168	8,343	14,68	21,67
10	2,558	4,865	9,342	15,99	23,21
11	3,053	5,578	10,34	17,28	24,73
12	3,571	6,304	11,34	18,55	26,22
13	4,107	7,042	12,34	19,81	27,69
14	4,660	7,790	13,34	21,06	29,14
15	5,229	8,547	14,34	22,31	30,58
20	8,260	12,44	19,34	28,41	37,57
25	11,52	16,47	24,34	34,38	44,31
30	14,95	20,60	29,34	40,26	50,89
40	22,16	29,05	39,34	51,81	63,69
50	29,71	37,69	49,34	63,17	76,15
60	37,49	46,46	59,34	74,40	88,38
70	45,44	55,33	69,33	85,53	100,4
80	53,54	64,28	79,33	96,58	112,3
90	61,75	73,29	89,33	107,6	124,1
100	70,07	82,36	99,33	118,5	135,8
$\infty$	$\nu - a$ $a = 3,290\sqrt{\nu}$	$\nu - b$	$\nu$	$\nu + b$ $b = 1,812\sqrt{\nu}$	$\nu + a$

## EENHEDEN

1

## DEFINITIES SI GRONDEENHEDEN

SI: *Système International d'Unités*

*lengte*: **meter** (m) afstand die licht in vacuum aflegt in 1/299 792 458 seconde (1983).

*massa*: **kilogram** (kg) massa van het internationale prototype van het kilogram (1901).

*tijd*: **seconde** (s) duur van 9 192 631 770 perioden van een hyperfijnovergang in cesium-133 atomen in de grondtoestand (1967).

*stroomsterkte*: **ampere** (A) stroomsterkte in twee oneindig lange en dunne parallelle geleiders, die op 1 m afstand een kracht van  $2 \times 10^{-7}$  N/m op elkaar uitoefenen (1948).

*thermodynamische temperatuur*: **kelvin** (K) fractie 1/273,16 van de thermodynamische temperatuur van het tripelpunt van water (1967).

*hoeveelheid stof*: **mol** (mol) hoeveelheid stof die evenveel elementaire entiteiten bevat als er atomen zijn in 0,012 kg zuiver koolstof-12. De entiteiten (atomen, moleculen, ionen, elektronen etc) moeten gespecificeerd zijn (1971).

*lichtsterkte*: **candela** (cd) lichtsterkte van een stralingsbron met frequentie  $540 \times 10^{12}$  Hz die in een gegeven richting 1/683 W/sr (watt per steradiaal) uitzendt (1979).

## Voorvoegsels voor S.I. eenheden:

$10^{-1}$	deci	d	$10^{-2}$	centi	c	$10^{-3}$	milli	m
$10^{-6}$	micro	$\mu$	$10^{-9}$	nano	n	$10^{-12}$	pico	p
$10^{-15}$	femto	f	$10^{-18}$	atto	a	$10^{-21}$	zepto	z
$10^{-24}$	yocto	y						
$10^1$	deka	da	$10^2$	hecto	h	$10^3$	kilo	k
$10^6$	mega	M	$10^9$	giga	G	$10^{12}$	tera	T
$10^{15}$	peta	P	$10^{18}$	exa	E	$10^{21}$	zetta	Z
$10^{24}$	yotta	Y						

Zie <http://physics.nist.gov/cuu/Units/>

NL000301.0711

© HB2009



## EENHEDEN

2

## AFGELEIDE SI EENHEDEN

vlakke hoek	$\alpha, \dots$	<b>radiaal</b>	rad ( $2\pi =$ cirkel)
ruimtehoek	$\omega, \Omega$	<b>steradiaal</b>	sr ( $4\pi =$ bol)
oppervlakte	$A, S$		$m^2$
inhoud	$V$		$m^3$
frequentie	$\nu$	<b>hertz</b>	Hz = $s^{-1}$
impuls	$p$		$kg\ m\ s^{-1}$
impulsmoment	$L, J$		$kg\ m^2\ s^{-1}$
soort. massa	$\rho$		$kg/m^3$
traagheidsmom.	$I$		$kg\ m^2$
kracht	$F$	<b>newton</b>	N = $kg\ m\ s^{-2}$
koppel	$M$		N m = $kg\ m^2\ s^{-2}$
druk	$p, P$	<b>pascal</b>	Pa = $N/m^2$
viskositeit	$\eta$		$N\ s\ m^{-2} =$ $kg\ m^{-1}\ s^{-1}$
energie	$E, w$	<b>joule</b>	J = N m = $kg\ m^2\ s^{-2}$
vermogen	$P$	<b>watt</b>	W = J/s
lading	$q, Q$	<b>coulomb</b>	C = A s
el. potentiaal	$V, \Phi$	<b>volt</b>	V = J/C
el. veldsterkte	$E$		V/m
diël. verplaatsing	$D$		C/m <sup>2</sup>
capaciteit	$C$	<b>farad</b>	F = C/V
weerstand	$R$	<b>ohm</b>	$\Omega = V/A$
specif. weerst.	$\rho$		$\Omega\ m$
geleiding	$G$	<b>siemens</b>	S = $\Omega^{-1}$
spec. gel.verm.	$\sigma, \kappa$		S/m
inductantie	$L$	<b>henry</b>	H = Wb/A
magn. flux	$\Phi$	<b>weber</b>	Wb = V s
magn. veldst.	$H$		A/m
magn. inductie	$B$	<b>tesla</b>	T = Wb/m <sup>2</sup>
lichtstroom	$\Phi$	<b>lumen</b>	lm = cd/sr
verlicht. st.	$I$	<b>lux</b>	lx = lm/m <sup>2</sup>
activiteit	$A$	<b>becquerel</b>	Bq = $s^{-1}$
geabsorb. dosis	$D$	<b>gray</b>	Gy = J/kg
dosisequivalent	$H$	<b>sievert</b>	Sv = J/kg

NL000302.0711

© HB2009

## EENHEDEN

3

**NIET-SI EENHEDEN (incl. Brit., US)** (zie ook **atomaire eenheden** op pag.5)

*lengte:* **fermi** (fm) =  $10^{-15}$  m; **Ångstrom** (Å) =  $10^{-10}$  m; **mil** (mil) = 0,001 in; **inch** (in) = 2,54 cm (exact); **foot** (ft) = 12 in = 0,304 8 m; **yard** (yd) = 3 ft = 0,914 4 m; **fathom (vadem)** = 6 ft = 1,828 8 m; **cable** = 720 ft = 185,2 m; **(statute) mile** = 1760 yd = 1609,34 m; **int. zeemijl** (nm) = 1852 m; **astronom. eenh.** (AE) = 1,495 978 70  $\times 10^{11}$  m; **parsec** (pc) =  $3,086 \times 10^{16}$  m.

*oppervlak:* **barn** (b) =  $10^{-28}$  m<sup>2</sup> = 100 fm<sup>2</sup>; **are** (a) = 100 m<sup>2</sup>; **hectare** (ha) = 10<sup>4</sup> m<sup>2</sup> (bunder); **acre** = 4840 sq. yd = 4046,87 m<sup>2</sup>; **sq. mile** = 640 acres = 2,59 km<sup>2</sup>.

*volume:* **br. fluid ounce** fl oz) = 28,41 cm<sup>3</sup>; **US fl. oz** = 29,572 9 cm<sup>3</sup>; **US liq. pint** = 16 US fl. oz = 473,2 cm<sup>3</sup>; **br. pint** (pt) = 20 br. fl. oz = 568,2 cm<sup>3</sup>; **US liq. quart** = 2 US liq. pt = 946,3 cm<sup>3</sup>; **liter** (l) = 1 dm<sup>3</sup>; **br. quart** (qt) = 2 br. pt = 1,136 dm<sup>3</sup>; **US gallon** = 4 US liq. qt = 231 in<sup>3</sup> = 3,785 4 dm<sup>3</sup>; **(br.) imperial gallon** (gal) = 4 br. qt = 4,546 dm<sup>3</sup>; **bushel** = 8 imp. gal; **barrel** = 42 US gal.; **ton** = 1 m<sup>3</sup>; **registerton** = 100 ft<sup>3</sup> = 2,83 m<sup>3</sup>.

*massa:* **u** (unified atomic mass unit) = 1,660 538 73(13)  $\times 10^{-27}$  kg; **grain avdp** (gr) = 64,79891 mg (exact); **(br.) drachme** = **(US) dram** = 60 gr = 3,887 934 6 g; **ounce avdp** (oz) = 28,349 527 g (exact); **troy ounce (apothecary ounce)** = 480 gr = 31,103 4768 g; **pound avoirdupois** (lb) = 16 oz = 7000 grain = 0,453 592 37 kg (exact); **(br.) stone** = 14 lbs = 6,35 kg; **ton** = 1000 kg.

*tijd:* **minuut** (min) = 60 s; **uur** (h) = 3600 s.

*temperatuur:* **t graad celsius** (°C) =  $t + 273,15$  K; **f graad Fahrenheit** (°F) =  $(f - 32) \times 5/9$  °C.

*snelheid:* **knoop** = zeemijl/h = 0,514 44 m/s.

*kracht:* **dyne** (dyn) =  $10^{-5}$  N; **poundforce** (lbf) = 4,448 22 N; **kilogramkracht** (kgf) = 9,806 65 N (ex.).

## EENHEDEN

4

*druk*: mm kwik (torr) = 101 325/760 Pa (exact) = 133,322 Pa; **pound per sq. inch** (psi) = 6 894,76 Pa; **technische atmosfeer** (at) = kgf/cm<sup>2</sup> = 98 066,5 Pa (exact); **bar** (bar) = 10<sup>5</sup> Pa; **normale atm.** = 101 325 Pa (exact).

*energie*: **hartree** ( $E_H$ ) = 4.359 744 17(75) × 10<sup>-18</sup> J; **erg** (erg) = 10<sup>-7</sup> J; **thermochemische calorie** (cal<sub>th</sub>) = 4,184 J; **15<sup>0</sup> calorie** (cal<sub>15</sub>) = 4,1855 J; **Int. Table calorie** (cal<sub>IT</sub>) = 4,1868 J; **br. thermal unit** (Btu) = 1055,87 J; **kilowattuur** (kWh) = 3,6 MJ; **ton steenkool equiv.** (tse) = 29,3 GJ; **ton olie equiv.** (toe) = 45,4 GJ; m<sup>3</sup> **aardgas** (gemiddeld, 0 °C, 1 atm) = 39,4 MJ.

*vermogen*: **paardekracht** (PK) = 75 kgf m/s = 735,5 W; **horsepower** (hp) = 550 lbf ft/s = 745,7 W.

*viskositeit*: **poise** (p) = g cm<sup>-1</sup>s<sup>-1</sup> = 0,1 kg m<sup>-1</sup>s<sup>-1</sup>; *kinematische viskositeit*: **stokes** (St) = 10<sup>-4</sup> m<sup>2</sup>/s

*(radio)activiteit, dosis*: **curie** (Ci) = 3,7 × 10<sup>10</sup> Bq; **röntgen** (R) = 2,58 × 10<sup>-4</sup> C/kg; **rad** (rad, rd) = 0,01 Gy; **rem** (rem) = 0,01 Sv.

*licht*: **stilb** (sb) = cd/cm<sup>2</sup>; **phot** (ph) = cd cm<sup>-2</sup>sr<sup>-1</sup>.

**elektrostatische eenheden (ese)**: c.g.s. eenh. van lading (g<sup>1/2</sup>cm<sup>3/2</sup>s<sup>-1</sup>) zodanig dat 4πϵ<sub>0</sub> = 1 (dimensieloos): *lading*: 10<sup>-9</sup>/2,997 924 58 C; *stroom*: 10<sup>-9</sup>/2,99... A; *dipoolmoment*: 10<sup>-11</sup>/2,99... C m; **debye** (D) = 10<sup>-18</sup> ese = 10<sup>-29</sup>/2,99... C; *el. pot.*: 299,7... V; *el. veldsterkte*: 2,99... × 10<sup>4</sup> V/m.

**elektromagnetische eenheden (eme)**: c.g.s. eenh. van stroom (g<sup>1/2</sup>cm<sup>1/2</sup>s<sup>-1</sup>) zodanig dat μ<sub>0</sub>/4π = 1 (dimensieloos): *Stroom*: **abampere** (abamp) = 10 A; *magneetveld*: **oerstedt** (Oe) = (1/4π) abamp/cm = 10<sup>3</sup>/4π A/m; *magn. inductie*: **gauss** (G) = 10<sup>-4</sup> T; *magn. flux*: **maxwell** (Mx) = 10<sup>-8</sup> Wb.

EENHEDEN		5
<b>ATOMAIRE EENHEDEN (a.u.)</b>		
Deze berusten op de Bohr straal $a_0$ , elektronmassa $m_e$ , $c$ van Dirac $\hbar$ en elektronlading $e$ .		
○	massa	$m_e = 9,109\,382\,15(45) \times 10^{-31}$ kg
	lengte	$a_0 = 5,291\,772\,0859(36) \times 10^{-11}$ m
	lading	$e = 1,602\,176\,487(40) \times 10^{-19}$ C
	tijd	$a_0/(ac) = 2,418\,884\,326\,505(16) \times 10^{-17}$ s
○	snelheid	$ac = 2,187\,691\,2541(15) \times 10^6$ m/s
	energie	$\hbar^2/(ma_0^2) = e^2/(4\pi\epsilon_0) = \alpha^2 mc^2$
	(hartree)	$E_H = 4,359\,743\,94(22) \times 10^{-18}$ J
		$= 2\,625,312\,93(13)$ kJ/mol
		$= 627,464\,850(32)$ kcal/mol
		$= 27,211\,383\,86(68)$ eV
○	$m_e = 1$ a.u., $\hbar = 1$ a.u., $c = 1/\alpha$ a.u., $e = 1$ a.u., $4\pi\epsilon_0 = 1$ a.u.	
<b>MD ('Molecular Dynamics') EENHEDEN</b>		
Dit is een consistent eenhedenstelsel voor 'moleculaire' grootheden, te gebruiken bij moleculaire modellering en dynamica. Coulombkrachten hebben een <i>elektrische factor</i> coëfficiënt $f = 1/(4\pi\epsilon_0)$ : $F = fq_1q_2/r^2$ (zie tabel). De eenheid van $f$ is $\text{kJ mol}^{-1} \text{nm e}^{-2}$ .		
○	massa	u $= 1,660\,538\,782(83) \times 10^{-27}$ kg
	lengte	nm $= 10^{-9}$ m
	tijd	ps $= 10^{-12}$ s
○	snelheid	nm/ps $= 1000$ m/s
	energie	kJ/mol $= 1,660\,538\,782 \times 10^{-21}$ J
	kracht	$\text{kJ mol}^{-1} \text{nm}^{-1} = 1,660\,538\,782 \times 10^{-12}$ N
	druk	$\text{kJ mol}^{-1} \text{nm}^{-3} = 1,660\,538\,782 \times 10^5$
		$= 16,605\,388$ bar
○	lading	$e = 1,602\,176\,487(40) \times 10^{-19}$ C
	el.factor	$f = 138,935\,460(7)$
NL000305.0711		© HB2009



FYSISCHE CONSTANTEN		1/2
<i>(tussen haakjes: standaardafwijking)</i>		
○	lichtsnelheid	$c = 299\,792\,458$ m/s (exact)
	magnet. const.	$\mu_0 = 4\pi \times 10^{-7}$ N/A <sup>2</sup> (exact) $= 1,256\,637\,0614\dots \times 10^{-6}$
	elektrische const.	$\varepsilon_0 = 1/\mu_0 c^2$ (exact) $= 8,854\,187\,817\dots \times 10^{-12}$ F/m
	karakt. imped. vacuum	$Z = \sqrt{\mu_0/\varepsilon_0} = \mu_0 c$ (exact) $= 376,730\,313\,461\dots \Omega$
○	Planck co.	$h = 6,626\,068\,96(33) \times 10^{-34}$ J s
	Dirac co. $\hbar/2\pi$	$\hbar = 1,054\,571\,628(53) \times 10^{-34}$ J s
	gravitatieconst.	$G = 6,674\,28(67) \times 10^{-11}$ m <sup>3</sup> kg <sup>-1</sup> s <sup>-2</sup>
	lading elektron	$e = 1,602\,176\,487(40) \times 10^{-19}$ C
	massa elektron	$m_e = 9,109\,382\,15(45) \times 10^{-31}$ kg
	massa proton	$m_p = 1,672\,621\,637(83) \times 10^{-27}$ kg $= 1,007\,276\,466\,77(10)$ u
○	$m_e/m_p$	$= 5,446\,170\,2177(24) \times 10^{-4}$
	atom. massaenh.	$u = 1,660\,538\,782(83) \times 10^{-27}$ kg
	Avogadro getal	$N_A = 6,022\,141\,79(30) \times 10^{23}$ mol <sup>-1</sup>
	Boltzmann const.	$k = 1,380\,6504(24) \times 10^{-23}$ J/K
	gasconstante $kN_A$	$R = 8,314\,472(15)$ J mol <sup>-1</sup> K <sup>-1</sup>
○	molair volume	$V_m = 22,710\,98(40) \times 10^{-3}$ m <sup>3</sup> /mol (ideaal gas 273,15 K, 100 kPa)
	Faraday $eN_A$	$F = 96\,485,3399(24)$ C/mol
	Bohr straal	$a_0 = 5,291\,772\,0859(36) \times 10^{-11}$ m $a_0 = \hbar/(m_e c \alpha) = 10^7 (\hbar/c e)^2/m_e$
	Bohr magneton	$\mu_B = 9,274\,009\,15(23) \times 10^{-24}$ J/T $\mu_B = e\hbar/2m_e$
○	nucl. magneton	$\mu_N = 5,050\,783\,24(13) \times 10^{-27}$ J/T
	magn.mom. el.	$\mu_e = -9,284\,763\,77(23) \times 10^{-24}$ J/T
	magn.mom. pr.	$\mu_p = 1,410\,606\,662(37) \times 10^{-26}$ J/T
	$g$ -factor elektron	$g_e = -2,002\,319\,304\,3622(15)$
	$g$ -factor proton	$g_p = 5,585\,694\,713(46)$
	fijnstructuurconst.	$\alpha = 7,297\,352\,5376(50) \times 10^{-3}$
○	$\alpha^{-1} = 4\pi\varepsilon_0\hbar c/e^2$	$\alpha^{-1} = 137,035\,999\,679(94)$
	gyromagn.verh.pr.	$\gamma_p = 2,675\,222\,099(70) \times 10^8$ s <sup>-1</sup> T <sup>-1</sup>
NL000101-0710		© HB2009

## FYSISCHE CONSTANTEN

2/2

geleidingsquantum  $G_0 = 7,748\,091\,7004(53) \times 10^{-5}$  S  
 Josephson const.  $K_J = 4,835\,978\,91(12) \times 10^{14}$  Hz/V  
 magn. flux quant.  $\Phi_0 = 2,067\,833\,667(52) \times 10^{-15}$  Wb  
 $G_0 = 2e^2/h$ ;  $K_J = 2e/h$ ;  $\Phi_0 = h/2e$   
 Stefan-Boltzm.co.  $\sigma = 5,670\,400(40) \times 10^{-8}$   
 $\pi^2 k^4 / (60 \hbar^3 c^2)$ ;  $U = \sigma T^4$  (zwarte str.)  $\text{W m}^{-2} \text{K}^{-4}$   
 Rydberg constant  $R_\infty = 10\,973\,731,568\,527(73) \text{ m}^{-1}$   
 $\alpha^2 m_e c / 2h$

massa's van neutron ( $n$ ), deuteron ( $d$ ) en muon ( $\mu$ )

$n$ :  $1,674\,927\,211(84) \times 10^{-27}$  kg =  $1,008\,664\,915\,97(43)$  u

$d$ :  $3,343\,583\,20(17) \times 10^{-27}$  kg =  $2,013\,553\,212\,724(78)$  u

$\mu$ :  $1,883\,531\,30(11) \times 10^{-28}$  kg =  $0,113\,428\,9256(29)$  u

**Relatieve standaarddeviaties**

$g_e$	$7,4 \times 10^{-13}$	$g_p$	$8,2 \times 10^{-9}$
$R_\infty$	$6,6 \times 10^{-12}$	$e, K_J, \Phi_0$	$2,5 \times 10^{-8}$
$m_d/u$	$3,9 \times 10^{-11}$	$h, N_A, u, m_e,$	
$m_p/u$	$1,0 \times 10^{-10}$	$m_p, m_d, m_n$	$5,0 \times 10^{-8}$
$m_e/u, m_n/u,$		$k, R, V_m$	$1,7 \times 10^{-6}$
$m_e/m_p$	$4,2 \times 10^{-10}$	$\sigma$ (Stefan-B.)	$7,0 \times 10^{-6}$
$\alpha, a_o, G_0$	$6,8 \times 10^{-10}$	$G$	$1,0 \times 10^{-4}$

**Nauwkeurigheden van afgeleide grootheden**

Stel grootheden  $x_i$ ,  $i = 1, \dots, N$ , zijn bekend met (co)varianties  $\sigma_{ij}$ . De (co)varianties  $\sigma_{kl}$  van afgeleide grootheden  $y_k$  zijn dan:

$$\sigma_{kl} = \sum_{i,j=1}^N \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \sigma_{ij}.$$

Als  $y_k$  uit machten van fysische constanten  $x_i$  is samengesteld:  $y_k = a_k \prod_{i=1}^N x_i^{p_{ki}}$  ( $a_k$  is een constante), dan is

$$\epsilon_k^2 = \sum_{i=1}^N p_{ki}^2 \epsilon_i^2 + 2 \sum_{j < i}^N p_{ki} p_{kj} r_{ij} \epsilon_i \epsilon_j,$$

waar  $\epsilon_k$  = relatieve standaarddeviatie en  $r_{ij}$  = correlatiecoëfficiënt tussen  $i$  en  $j$  (Voor  $r$ : zie website)

CODATA 2006 <http://physics.nist.gov/cuu/constants/>

NL000102-0710

© HB2009

## KL. KWADR. AANPASSING

1

**ALGEMEEN**

Gegeven  $n$  meetwaarden  $y_i$ ,  $i = 1, \dots, n$ , gevraagd  $m$  parameters  $\theta_k$ ,  $k = 1 \dots m$ ,  $m < n$ ; zodat:

$$\chi^2 = \sum_{i=1}^n (y_i - f_i)^2 / \sigma_i^2 \text{ minimaal.}$$

$f_i(\theta_1, \dots, \theta_m)$  zijn de functiewaarden die op  $y_i$  aangepast worden. Zowel  $y_i$  als  $f_i$  kunnen functies van onafhankelijke variabelen zijn. De afwijkingen  $(y_i - f_i)$  worden geacht onafhankelijke stochastische grootheden te zijn met verwachtingswaarde 0 en bekende variantie  $\sigma_i^2$ . De gewichtsfactor van de  $i^e$  meetwaarde is  $1/\sigma_i^2$ . De som van gewogen kwadratische afwijkingen  $\chi^2$  moet bij ideale aanpassing voldoen aan een  $\chi^2$ -verdeling voor  $n - m$  vrijheidsgraden.

**LINEAIR IN DE PARAMETERS**

Stel  $f_i$  is een lineaire functie van  $\theta_1, \dots, \theta_m$ :

$$f_i(\theta) = \sum_k a_{ik} \theta_k; \quad a_{ik} = \partial f_i / \partial \theta_k; \quad \mathbf{f} = \mathbf{A} \boldsymbol{\theta}$$

De kleinste-kwadratenoplossing is dan:

$$\boldsymbol{\theta} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{y} \quad \text{waar } \mathbf{W}_{ij} = \sigma_i^{-2} \delta_{ij}$$

Verwachtingswaarden van covarianties van  $\boldsymbol{\theta}$ :

$$E\{\Delta \theta_k \Delta \theta_l\} = [(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1}]_{kl} \chi^2 / (n - m)$$

**BIJZONDER GEVAL:**  $f_i = f(x_i) = ax_i + b$

$$a = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2}; \quad b = \langle y \rangle - a \langle x \rangle$$

Verwachtingswaarden van (co)varianties van  $a$  en  $b$ :

$$E\{(\Delta a)^2\} = \sigma_a^2 = \chi^2 / [n(n-2)(\langle x^2 \rangle - \langle x \rangle^2)]$$

$$E\{(\Delta b)^2\} = \sigma_b^2 = \langle x^2 \rangle \sigma_a^2; \quad E\{\Delta a \Delta b\} = -\langle x \rangle \sigma_a^2$$

N.B.:  $a$  en  $b$  zijn onafhankelijk als  $\langle x \rangle = 0$

Correlatiecoefficient  $r$ :

$$r = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sqrt{(\langle x^2 \rangle - \langle x \rangle^2)} \sqrt{(\langle y^2 \rangle - \langle y \rangle^2)}} = a \left( \frac{\langle x^2 \rangle - \langle x \rangle^2}{\langle y^2 \rangle - \langle y \rangle^2} \right)^{1/2}$$

Hier zijn  $\langle \rangle$  de met  $\sigma^{-2}$  gewogen gemiddelden:

$$\langle \xi_i \rangle = [\sum_{i=1}^n \xi_i / \sigma_i^2] / [\sum_{i=1}^n 1 / \sigma_i^2]$$





MATHEMATICA

1

(Summiere samenvatting; *z* handboek en **help** functie)

**Shift-Enter**: voer opdracht uit.

**Symbolen, getallen, strings** (willekeurige lengte)

<i>xxx</i>	globaal gebruikerssymbool
<i>Xxxx</i> , <i>\$Xxxx</i>	globaal systeemssymbool
<i>xxx\$</i> , <i>xxx\$nn</i>	lokaal symbool, in <i>module</i>
$[\pm]nn$	<b>Integer</b>
$[\pm]nn.nn$	<b>Real</b>
$[\pm]nn.nn 10^{[\pm]nn}$	idem
$bb^{mm[.mm]}$	<i>bb</i> = basis 2-36, <i>mm</i> = 0-9, a-z
<i>nn/nn</i>	<b>Rational</b>
<i>nn[.nn] + nn[.nn] I</i>	<b>Complex</b>
" <i>ccc</i> "	string ASCII, 8- of 16-bit tekens
$\backslash nnn$ , $\backslash .mm$	8-bit teken ( <i>nnn</i> oct, <i>mm</i> hexadec)
$\backslash ^A.. \backslash ^Z$ , $\backslash n$ , $\backslash f$	Ctrl-A .. Ctrl-Z, newline, newpage
$\backslash "$ , $\backslash \backslash$ , $\backslash b$ , $\backslash t$ , $\backslash r$	" , \, backspace, tab, return

**Gebruik van haakjes**

(...)	groepering	$[x, y]$	functie van
{ <i>x, y</i> }	lijst van objecten	$x[[n]]$	$n^e$ el. van lijst <i>x</i>
(* <i>any text</i> *)	commentaar		

**Constanten**

Pi	$\pi = 3.1415\dots$	Infinity	$\infty$
I	$\sqrt{-1}$	GoldenRatio	$(1 + \sqrt{5})/2$
E	$e = 2.718\dots$	EulerGamma	$\gamma = 0.577\dots$
Degree	$\pi/180$	Catalan	0.915966...

**Random getallen**

Random[]	random getal tussen 0. en 1.
Random[ <i>type</i> ]	random getal van <i>type</i> =
Random[ <i>type, range</i> ]	<b>Integer, Real of Complex</b>
Random[ <i>type, range, n</i> ]	in <i>range</i> (default 0 tot 1) met <i>n</i> digits precisie
<i>range</i> =	{ <i>min, max</i> }, default {0, <i>range</i> }
SeedRandom[]	seed = time-of-day
SeedRandom[ <i>n</i> ]	seed = integer <i>n</i> .

## MATHEMATICA

2

**Edit bewerkingen**

; scheidt commandos en  
     onderdrukt output  
 %% laatste (voorlaatste) resultaat  
 %n of Out[n] resultaat van *Out[n]*  
 In[n] (In[]) herhaal  $n^e$  (laatste) input  
 InString[n] tekst van  $n^e$  input  
 InString[] idem laatste input  
 ?*Naam* geef informatie over *Naam*  
 ??*Naam* meer informatie over *Naam*  
 Information[*Naam*] idem  
 ?A\* alle namen die met A beginnen

**Definities**

*lhs* = *rhs* *rhs* wordt direct geëvalueerd  
*lhs* := *rhs* *rhs* wordt niet geëvalueerd  
*x* = *y* = 2.5 meervoudige toewijzing toegestaan  
*f*[*x\_*] = *expr* definieert *f* als functie van *x*  
*x* = . of Clear[*x*] maakt toewijzing aan *x* ongedaan

**Definities van lijsten en matrices**

{*x*<sub>1</sub>, *x*<sub>2</sub>, ...} lijst, array, vector  
 {{*x*<sub>11</sub>, *x*<sub>12</sub>, ...} {*x*<sub>21</sub>, *x*<sub>22</sub>, ...} ...} matrix *x*<sub>*ij*</sub>  
 Hieronder: range = *imax*; *i*, *imax*; (van 1 tot *imax*)  
                   *i*, *imin*, *imax*; *i*, *imin*, *imax*, *di*  
 List[*x*<sub>1</sub>, *x*<sub>2</sub>, ...] maakt lijst  
 Table[*expr*, {range}] geeft lijst waarden van *expr*  
 Array[*x*, *n*] geeft {*x*[1], ..., *x*[*n*]}  
 Range[*n*] geeft {1, 2, ..., *n*}  
 IdentityMatrix[*n*] geeft  $n \times n$  eenheidsmatrix  
 DiagonalMatrix[*lijst*] geeft matrix met *lijst* op diag.

**Logische operatoren (True of False)**

<i>x</i> == <i>y</i>	gelijk		<i>x</i> != <i>y</i>	ongelijk
<i>x</i> > <i>y</i>	groter dan		<i>x</i> >= <i>y</i>	groter of gelijk
<i>x</i> < <i>y</i>	kleiner dan		<i>x</i> <= <i>y</i>	kleiner of gelijk
! <i>p</i>	niet (negatie) <i>p</i>		<i>p</i> && <i>q</i>	en
<i>p</i>    <i>q</i>	of		Xor[ <i>p</i> , <i>q</i> ]	exclusive or

## MATHEMATICA

3

**Numerieke bewerkingen**

$\dots/N$  of  $N[\dots]$  geeft numerieke waarde.

$N[\dots, n]$  numerieke waarde met  $n$  cijfers.

**Symbolische bewerkingen**

$-x$  minus |  $x^y$  macht

$x+y$  som |  $x-y$  verschil

$x*y$  of  $x y$  product |  $x/y$  quotient

z. ook: Simplify, Expand, Factor

**Voorwaardelijke bewerkingen**

$lhs := rhs /; test$  definiëer alleen als  
 $test$  True is.

If[ $cond, t, f$ ]  $t$  als  $cond$  True is,

If[ $cond, t, f, u$ ]  $f$  als  $cond$  False is,  
anders  $u$ .

Switch[ $expr, x_1, f_1, x_2, \dots$ ] geeft  $f_i$  voor eerste  $x_i$   
die gelijk is aan  $expr$ .

Which[ $c_1, f_1, c_2, \dots$ ] geeft  $f_i$  voor eerste  $c_i$   
die True is.

**Herhaalde bewerkingen**

Do[ $expr, \{imax\}$ ] Voer  $expr$  uit voor  $i$  van

Do[ $expr, \{i, imin, imax\}$ ]  $imin$  (default 1) tot

Do[ $expr, \{i, imin, imax, di\}$ ]  $imax$  met stappen  
 $di$  (default 1).

FixedPoint[ $f, expr$ ] Pas  $f$  herhaald toe op  
 $expr$  tot resultaat niet  
meer verandert.

FixedPoint[ $f, expr,$   
SameTest  $\rightarrow comp$ ] idem tot  $comp$  toegepast  
op opvolgende  
resultaten True is.

For[ $start, test, incr, body$ ] evalueer  $start$ , voer dan  
For[ $i=1, i<=10, i++, \dots$ ]  $body$  en  $incr$  uit  
zolang  $test$  is True.

Nest[ $f, expr, n$ ] pas functie  $f$   $n$  keer toe.  
op  $expr$

While[ $test, body$ ] voer  $body$  uit zolang  
 $test$  True is.

NL003003.9607

© HB2009

<b>MATHEMATICA</b>	4
--------------------	---

**Controlecommando's in lussen:**

<b>Break</b> []	verlaat de huidige lus.	○
<b>Continue</b> []	ga naar volgende stap in huidige lus.	○
<b>Return</b> [ <i>expr</i> ]	verlaat procedure met waarde <i>expr</i> .	○
<b>Goto</b> [ <i>naam</i> ]	ga naar <b>Label</b> [ <i>naam</i> ].	○

**Funkties**

**Abs**[*z*]; **Arg**[*z*]; **Cos**[*z*]; **Cot**[*z*]; **Csc**[*z*]; **Exp**[*z*]; **Im**[*z*];  
**Log**[*z*]; **Log**[*b, z*]; **Mod**[*n, m*]; *n!*; **Round**[*x*]; **Sec**[*z*];  
**Sin**[*z*]; **Sqrt**[*z*]; **Tan**[*z*]

*Ook ArcSin, Sinh, ArcSinh etc.*

*z. help:* **Bessel**, **Beta**, **Chebyshev**, **CosIntegral**,  
**Elliptic**, **Erf**, **Erfc**, **ExpIntegral**, **Gamma**,  
**Hypergeometric**, **LaguerreL**, **LegendreP**, **SinIntegral**

**Bewerkingen van lijsten en matrices**

Normale bewerkingen en functies werken elementsgewijs

**Sort**[*lijst*] sorteer lijst

**Length**[*lijst*] geeft lengte (aantal elementen) van lijst

**Min**, **Max**[*lijst*] minimum of maximum element uit lijst

**Apply**[**Plus**, *lijst*] geeft som van elementen (ook: **Times**)

**Fourier**[*lijst*], **InverseFourier**[*lijst*] complexe FT

**Transpose**[*matrix*] transposeer matrix

**Det**[*matrix*] determinant van vierkante matrix

**Inverse**[*matrix*] inverse van vierkante matrix

**Eigenvalues**[*matrix*] alleen eigenwaarden

**Eigenvectors**[*matrix*] alleen eigenvectoren

**Eigensystem**[*matrix*] → {eigenwaarden, vectoren}

**MatrixExp**[*matrix*], **MatrixPower**[*matrix, n*]

**Dot**[*a, b*] of *a.b* vector en matrixproducten

**LinearSolve**[*m, b*] geeft *x* als oplossing van  $m \cdot x = b$

**SingularValues**[*M*] geeft singuliere waarde-decompositie van matrix  $M = U^T W V$ , waar *U* en *V* orthonormale rijen hebben en *W* diagonaal is met de singuliere waarden.

Output is {*U, W<sub>ii</sub>, V*}.

**PseudoInverse**[*M*] geeft  $M' = V^T W^{-1} U$  (*M'* is kl. kwadr. opl. van  $MM' = I$ )

NL003004.9607	© HB2009
---------------	----------

## MATHEMATICA

5

**Bewerkingen op functies**

- `Derivative`[ $n_1, n_2, \dots$ ][ $f$ ] differentieer  $f$   $n_1$  keer naar eerste argument, etc;  $f' = \text{Derivative}[1][f]$
- `D`[ $f, x$ ] partiële afgeleide van  $f$  naar  $x$
- `Integrate`[ $f, x$ ] onbepaalde integraal
- `Integrate`[ $f, \{\text{range}\}$ ] bepaalde integraal; `NIntegrate`
- `InverseFunction`[ $f$ ] inverse functie van  $f$
- `Series`[ $\text{expr}, x, x_0, n$ ] machtreeks rond  $x_0$  t/m orde  $n$
- `DSolve`[ $\text{eqn}, y[x], x$ ] los differentiaalvlg op; `NDSolve`
- `Sum`, `NSum`, `Product`, `NProduct`[ $f, \{\text{range}\}$ ], `Solve`, `NSolve`, `FindMinimum`, `FindRoot`, `Limit`, `Residue`

**Data input en output**

- Inlezen van data op *file*: (!!*file* bekijk inhoud)
- `ReadList`["*file*", *Number*] maakt lijst van getallen
- `ReadList`["*file*", {*Number*, *Number*}] array van 2 kol.
- `ReadList`["*file*", *Table*[*Number*, {*n*}]] array van  $n$  kol.
- `ReadList`["*file*", *Number*, *RecordLists* -> *True*] geeft lijst per regel.
- I.p.v. *Number* ook toegestaan: *Byte*, *Character*, *Real*, *Word*, *Record*, *String*, *Expression*
- Schrijf tabel (*matrix*), tekst of figuur (*ps*) naar *file*:  
 $f = \text{OpenWrite}$ ["*file*", *FormatType* -> *OutputForm*];  
 $\text{Write}$ [ $f, \text{TableForm}[\text{matrix}]$ ];  $\text{WriteString}$ [ $f, \dots$ ];  
 $\text{Display}$ [ $f, \text{graphics}$ ] maakt postscript file van *graphics*, te bewerken met *psfix*; eindig met `Close`[ $f$ ].

**Grafische commando's**

- `ListPlot`[{ $y_1, y_2, \dots$ }] plot  $y$ , met  $x = 1, 2, \dots$
- `ListPlot`[{{ $x, y_1$ }, { $x, y_2$ }, ...}] plot waarden van  $x, y$
- `ListPlot3D`[*array*] geeft 3D-plot van hoogtewaarden, ook: `ListContourPlot`[*array*] en `ListDensityPlot`[*array*]
- `Plot`[ $f[x], \{x, x_{\min}, x_{\max}\}$ ] plot  $f(x)$  in *range*
- `Plot`[{ $f[x], g[x], \dots$ }, { $x, x_{\min}, x_{\max}$ }]
- `Plot3D`[ $f[x, y], \{x, x_{\min}, x_{\max}\}, \{y, y_{\min}, y_{\max}\}$ ], ook: `ContourPlot`[...], `DensityPlot`[...]
- `Show`[ $g_1, g_2, \dots$ ] laat grafieken  $g_1, g_2, \dots$  zien.

## MATHEMATICA

6

**Grafische opties**

*optie* → *waarde* kan in grafische commando's worden opgenomen, of apart gedeclareerd met `SetOptions[s,...]`, waar *s* het commando is waarvoor de opties moeten gelden (bv `Plot`). Opties: `Automatic`, `None`, `All`, `True`, `False`

<i>optie</i>	<i>default</i>	<i>waarde</i>
<code>AspectRatio</code>	<code>1/GoldenRatio</code>	hoogte/breedte
<code>Axes</code>	<code>True</code>	bv. <code>{False, True}</code>
<code>AxesLabel</code>	<code>None</code>	<code>"ylabelöf { "xlabel", "ylabel" }</code>
<code>AxesOrigin</code>	<code>Automatic</code>	<code>{x, y}</code>
<code>Compiled</code>	<code>True</code>	compile function
<code>Frame</code>	<code>False</code>	
<code>FrameLabel</code>	<code>None</code>	<code>{lowx, lefty, upx, righty}</code>
<code>FrameTicks</code>	<code>Automatic</code>	
<code>GridLines</code>	<code>None</code>	<code>Automatic</code>
<code>MaxBend</code>	<code>10</code>	max kink angle
<code>PlotDivision</code>	<code>20</code>	max subdiv. interval
<code>PlotJoined</code>	<code>False</code>	<code>True</code> : join points in <code>ListPlot</code>
<code>PlotLabel</code>	<code>None</code>	<code>"text"</code>
<code>PlotPoints</code>	<code>25</code>	nr of sample points
<code>PlotRange</code>	<code>Automatic</code>	<code>{yrange}</code> ( <i>z</i> for 3D), of <code>{xrange, yrange}</code>
<code>PlotStyle</code>	<code>Automatic</code>	zie onder

**PlotStyle opties**

`PlotStyle` → *style* of `{{style1}, {style2}, ...}` (de laatste worden cyclisch gerealiseerd): lijn- en achtergrondstijl  
`GrayLevel[c]` grijsniveau,  $0 \leq c \leq 1$ , wit = 1  
`RGBColor[r, g, b]` kleurcomponenten tussen 0 en 1  
`Hue[h]` of `Hue[h, s, b]` tint, verzadiging, helderheid (0, 1).  
`Thickness[t]` lijndikte *t* in fractie van totale breedte  
`Dashing[{l1.l2, ...}]` streppatroon (cyclisch)  
`Background` → *color*: achtergrondkleur;  
*color* = `GrayLevel`, `RGBColor`, `Hue`.

NL003006.9607

© HB2009

**NORMALE VERDELING**

1

**EENDIMENSIONALE GAUSSFUNCTIE**

$$f(x) dx = (\sigma\sqrt{2\pi})^{-1} \exp[-(x - \mu)^2/(2\sigma^2)] dx$$

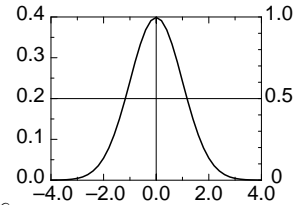
standaardvorm:

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

$\mu$  = gemiddelde

$\sigma^2$  = variantie

$\sigma$  = standaarddeviatie



**Centrale momenten**  $\mu_n = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$

$\mu_m = 0$  voor  $m$  even;  $\mu_{2n} = \sigma^{2n} \times 1 \times 3 \times 5 \times \dots \times (2n - 1)$

$\mu_2 = \sigma^2$ ;  $\mu_4 = 3\sigma^4$ ;  $\mu_6 = 15\sigma^6$ ;  $\mu_8 = 105\sigma^8$

$skewness = 0$ ;  $kurtosis = 0$

**Cumulatieve verdelingsfunctie:**

$$F(x) = \Pr\{x' < x\} = \int_{-\infty}^x f(x') dx'$$

$$1 - F(x) = F(-x) = \Pr\{x' > x\} = \int_x^{\infty} f(x') dx'$$

$$F(x) = \frac{1}{2} \{1 + \text{erf}(x/\sigma\sqrt{2})\}$$

$$1 - F(x) = \frac{1}{2} \{1 - \text{erf}(x/\sigma\sqrt{2})\} = \frac{1}{2} \text{erfc}(x/\sigma\sqrt{2})$$

$z$	$f(z)$	$F(-z)$	$z$	$f(z)$	$F(-z)$
0.0	0.3989	0.5000	1.4	1.497e-01	8.076e-02
0.1	0.3970	0.4602	1.6	1.109e-01	5.480e-02
0.2	0.3910	0.4207	1.8	7.895e-02	3.593e-02
0.3	0.3814	0.3821	2.0	5.399e-02	2.275e-02
0.4	0.3683	0.3446	2.5	1.753e-02	6.210e-03
0.5	0.3521	0.3085	3.0	4.432e-03	1.350e-03
0.6	0.3332	0.2743	3.5	8.727e-04	2.326e-04
0.7	0.3123	0.2420	4.0	1.338e-04	3.167e-05
0.8	0.2897	0.2119	5.0	1.487e-06	2.866e-07
0.9	0.2661	0.1841	7.0	9.135e-12	1.280e-12
1.0	0.2420	0.1587	10	7.695e-23	7.620e-24
1.2	0.1942	0.1151	15	5.531e-50	3.671e-51

grote  $z$  :  $F(-z) = 1 - F(z) = \frac{f(z)}{z} \left(1 - \frac{1}{z^2+2} + \dots\right)$



## NORMALE VERDELING

2

**Overschrijdingskansen**

Kans dat minstens één uit  $n$  (onafhankelijke, normaal verdeelde) steekproeven buiten het interval  $(\mu - d, \mu + d)$  valt:

$$\Pr\{\geq 1; n, d\} = 1 - [1 - 2F(-d/\sigma)]^n$$

$d/\sigma$	1,5	2	2,5	3	3,5	4
$n \downarrow$						
1	0,134	0,046	0,012	0,0027	4,7e-4	6,3e-5
2	0,249	0,089	0,025	0,0054	9,3e-4	1,3e-4
3	0,350	0,130	0,037	0,0081	0,0014	1,9e-4
4	0,437	0,170	0,049	0,0108	0,0019	2,5e-4
5	0,512	0,208	0,061	0,0134	0,0023	3,2e-4
6	0,577	0,244	0,072	0,0161	0,0028	3,8e-4
7	0,634	0,278	0,084	0,0187	0,0033	4,4e-4
8	0,683	0,311	0,095	0,0214	0,0037	5,1e-4
9	0,725	0,342	0,106	0,0240	0,0042	5,7e-4
10	0,762	0,372	0,117	0,0267	0,0046	6,3e-4
12	0,821	0,428	0,139	0,0319	0,0056	7,6e-4
15	0,884	0,503	0,171	0,0397	0,0070	9,5e-4
20	0,943	0,606	0,221	0,0526	0,0093	0,0013
25	0,972	0,688	0,268	0,0654	0,0116	0,0016
30	0,986	0,753	0,313	0,0779	0,0139	0,0019
40	0,997	0,845	0,393	0,102	0,0184	0,0025
50	0,999	0,903	0,465	0,126	0,0230	0,0032
70	1,000	0,962	0,583	0,172	0,0321	0,0044
100	1,000	0,991	0,713	0,237	0,0455	0,0063
150	1,000	0,999	0,847	0,333	0,0674	0,0095
200	1,000	1,000	0,918	0,418	0,0889	0,0126
300	1,000	1,000	0,976	0,556	0,130	0,0188
400	1,000	1,000	0,993	0,661	0,167	0,0250
500	1,000	1,000	0,998	0,741	0,208	0,0312

strepen markeren 5% niveau

## STUDENT'S T-VERDELING

1

## STUDENT'S T-VERDELING

- Stel  $X$  is een normaal gedistribueerde variabele met verwachtingswaarde 0 en variantie  $\sigma^2$ , en  $Y^2/\sigma^2$  is een onafhankelijke chikwadraat-verdeelde variabele met  $\nu =$  vrijheidsgraden, dan is  $t = \frac{X\sqrt{\nu}}{Y}$  verdeeld volgens een Student's t-distributie  $f(t|\nu)$  met  $\nu$  vrijheidsgraden, onafhankelijk van  $\sigma$ :

○ 
$$f(t|\nu) dt = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} dt$$

**Toepassing: nauwkeurigheid van gemiddelde**

Stel  $x_1, \dots, x_n$  zijn  $n$  onafhankelijke samples uit een normale verdeling met onbekende verwachtingswaarde  $\mu$  en onbekende variantie  $\sigma^2$ ;

- $\langle x \rangle = \frac{1}{n} \sum_{i=1}^n x_i$ ;  $S = \sum_{i=1}^n (x_i - \langle x \rangle)^2$  en  $\hat{\sigma} = \sqrt{S/(n-1)}$ , dan is  $t = [(\langle x \rangle - \mu)\sqrt{n}]/\hat{\sigma}$  verdeeld volgens een Student's t-verdeling met  $\nu = n - 1$  vrijheidsgraden. De beste schatting voor  $\sigma$  is  $\hat{\sigma}$ . Als  $\sigma$  wel bekend is, dan is  $\langle x \rangle$  normaal verdeeld met verwachtingswaarde  $\mu$  en variantie  $\sigma^2/n$ . In dat geval voldoet  $\chi^2 = S/\sigma^2$  aan een chikwadraat-verdeling met  $\nu = n - 1$  vrijheidsgraden.

**Eigenschappen en momenten**

$f$  is symmetrisch:  $f(-t) = f(t)$ ; gemiddelde = 0

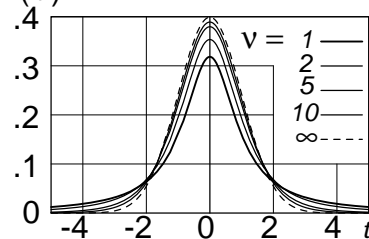
variantie  $\sigma^2 = \nu/(\nu - 2)$  ( $\nu > 2$ ); 'skewness'  $\gamma_1 = 0$

'excess'  $\gamma_2 = E\{t^4\}/\sigma^4 - 3 = 6/(\nu - 4)$

$\lim_{\nu \rightarrow \infty} f(t|\nu) = (1/\sqrt{2\pi}) \exp(-t^2/2)$

- 

$f(t)$

**CUMULATIEVE VERDELING**

$$F(t|\nu) = \int_{-\infty}^t f(t'|\nu) dt'$$

$$F(-t|\nu) = 1 - F(t|\nu)$$

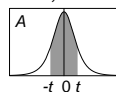
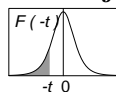
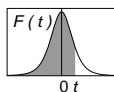
zie tabel p. 2

-

## STUDENT'S T-VERDELING

2

Waarden van  $t$  bij 75%, 90%, 95%, 99% en 99,5%



$A$  = acceptatieniveau voor tweezijdig interval  $(-t, t)$

$F(t) =$	0,75	0,90	0,95	0,99	0,995
$F(-t) =$	0,25	0,10	0,05	0,01	0,005
$A(\%)$	50	80	90	98	99

$\nu = 1$	1,000	3,078	6,314	31,821	63,657
2	0,816	1,886	2,920	6,965	9,925
3	0,765	1,638	2,353	4,541	5,841
4	0,741	1,533	2,132	3,747	4,604
5	0,727	1,467	2,015	3,365	4,032
6	0,718	1,440	1,943	3,143	3,707
7	0,711	1,415	1,895	2,998	3,499
8	0,706	1,397	1,860	2,896	3,355
9	0,703	1,383	1,833	2,821	3,250
10	0,700	1,372	1,812	2,764	3,169
11	0,697	1,363	1,796	2,718	3,106
12	0,695	1,356	1,782	2,681	3,055
13	0,694	1,350	1,771	2,650	3,012
14	0,692	1,345	1,761	2,624	2,977
15	0,691	1,341	1,753	2,602	2,947
20	0,687	1,325	1,725	2,528	2,845
25	0,684	1,316	1,708	2,485	2,787
30	0,683	1,310	1,697	2,457	2,750
40	0,681	1,303	1,684	2,423	2,704
50	0,679	1,299	1,676	2,403	2,678
60	0,679	1,296	1,671	2,390	2,660
70	0,678	1,294	1,667	2,381	2,648
80	0,678	1,292	1,664	2,374	2,639
90	0,677	1,291	1,662	2,369	2,632
100	0,677	1,290	1,660	2,364	2,626
$\infty$	0,674	1,282	1,645	2,326	2,576

# Index

**A**

absolute fouten 13  
 acre-definitie 119  
 AE-definitie 119  
 afrondfouten 32  
 ampere - definitie 117  
 angstrom-definitie 119  
 are-definitie 119  
 atmosfeer-definitie 120  
 atto - voorvoegsel 117  
 autocorrelatiefunctie 71  
 avdp-definitie 119  
 avogadrogetal 123

**B**

barn-definitie 119  
 barrel-definitie 119  
 becquerel-definitie 118  
**bestfit** 66  
 betrouwbaarheidsgrens 19  
 betrouwbaarheidsinterval 19,35  
 bias 83  
 binomiaalcoëfficiënt 21  
 binomiaalverdeling 20,85  
 block average 96  
 blokgemiddelde 96  
 bohr magneton 123  
 Boltzmann - const. van 123  
 bootstrap 42  
 btu-definitie 120

**C**

calorie-definitie 119  
 calorie-definitie 120  
 candela - definitie 117  
 celsius-definitie 119  
 centi - voorvoegsel 117  
 centrale limietstelling 22  
 centrale momenten 19  
 chi-kwadraat 101  
 chi-kwadraattest 68  
 chi-kwadratverdeling 24,115  
 chikwadratverdeling-cumulatief 116  
 confidence limits 19

constanten - fysische 123  
 contourplot 72  
 correlatiecoëfficiënt 72,79  
 correlatiecoëfficiënt-parameters 60  
 correlatielengte 34,59,92  
 coulomb-definitie 118  
 covariantie 60,79  
 covariantiematrix 66,72,100,101  
 covariantiematrix van meetwaarden 100

covariantiematrix van parameters 101

cumulatieve verdeling 19,29  
 curie-definitie 120

**D**

debye-definitie 120  
 deci - voorvoegsel 117  
 decimaalkomma 8  
 deka - voorvoegsel 117  
 deutron - massa 124  
 dirac - const. van 123  
 discrete kansverdeling 17  
 distributie-vrije methoden 41  
 doorwerking van fouten 14  
 drachme-definitie 119  
 dyne-definitie 119

**E**

Eadie-Hofstee plot 50  
 eenheden 9,117  
 elektron - g-factor 123  
 elektron - lading 123  
 elektron - massa 123  
 enzymkinetiek 50  
 erg-definitie 120  
 errorbars 46  
 ese-definitie 120  
 exa - voorvoegsel 117  
 excess 19  
 excess-schatting 36

**F**

fahrenheit-definitie 119  
 farad-definitie 118  
 faraday - constante 123

- fathom-definitie 119  
 femto - voorvoegsel 117  
 fermi-definitie 119  
 fijnstructuurconstante 123  
**Fit** 64  
 fl.oz.-definitie 119  
 foot-definitie 119  
 fout-middelbaar 19,34  
 fout-standaard 19  
 fouten afronden 9  
 fouten-absoluut 13  
 fouten-doorwerking 14  
 fouten-optellen 79  
 fouten-relatief 13  
 fouten-samengesteld 14  
 fouten-systematisch 12,83  
 fouten-toevallig 12  
 fysische constanten 123
- G**  
 g-factor - elektron 123  
 g-factor - proton 123  
 gallon-definitie 119  
 gasconstante 123  
 gauss-definitie 120  
 Gaussfunctie 22  
 gemiddelde 18  
 gemiddelde kwadratische afwijking 32
- gemiddelde van meetserie 30  
 gemiddelde-schatting 33  
 gemiddelde-standaarddeviatie van 95  
 getal van Avogadro 123  
 gewichten-ongelijk 37  
 gewichtsfactor 38  
 gewogen middeling 38  
 giga - voorvoegsel 117  
 grafische schattingen 52  
 grain-definitie 119  
 gravitatieconstante 123  
 gray-definitie 118  
 grondeenheden 117  
 gyromagnetische verhouding 123
- H**
- Hanes plot 50  
 hecto - voorvoegsel 117  
 henry-definitie 118  
 histogram 28
- I**  
 ijkfunctie 56  
 ijklijn 54  
 ijktabel 56  
 inch-definitie 119
- J**  
 jackknife procedure 96  
 joule-definitie 118
- K**  
 kWH-definitie 120  
 kansverdeling 18  
 kansverdeling-continu 17  
 kansverdeling-discreet 17  
 kelvin - definitie 117  
 kgkracht-definitie 119  
 kilo - voorvoegsel 117  
 kilogram - definitie 117  
 klassieke methoden 41  
 kleinste kwadraten-lineair 58  
 kleinste kwadraten - zakb. 125  
 kleinste kwadraten-algemeen 65  
 kleinste kwadraten-fout in  $x$  58  
 kleinste kwadraten-nauwkeurigheid 71
- kleinste kwadraten-niet lineair 66,103  
 kleinste kwadraten-varianties 59  
 kleinste kwadratenaanpassing 57,99  
 kleinste-kwadratenmethode 46  
 kurtosis 19  
 kwartiel 20
- L**  
 lineaire regressie 57,99  
 linearisatie 47  
 Lineweaver-Burk plot 50  
 liter-definitie 119  
 lumen-definitie 118  
 lux-definitie 118
- M**

- marginale verdeling 90
- mathematica - zakb. 127
- maxwell-definitie 120
- mean 18
- mediaan 20
- mega - voorvoegsel 117
- meter - definitie 117
- Michaelis-Menten kinetiek 50
- micro - voorvoegsel 117
- micron 10
- middelbare fout 9,19
- mil-definitie 119
- mile-definitie 119
- milli - voorvoegsel 117
- minimum-variantieschatting 97
- mm kwik-definitie 120
- mol - definitie 117
- moment van verdeling 19
- multinomiaalverdeling 24,85
- muon - massa 124
- N**
- nauwkeurigheden fys. const. 124
- neutron - massa 124
- newton-definitie 118
- normale verdeling 22
- normale verdeling - cumulatief 133
- normale verdeling - momenten 133
- normale verdeling - overschrijdingskans 134
- normale verdeling - zakb. 133
- normering 18
- notatie eenheden 9
- nucleair magneton 123
- O**
- oerstedt-definitie 120
- ohm-definitie 118
- ounce-definitie 119
- outliers 41
- overschrijdingskans 41
- P**
- paardekracht-definitie 120
- parsec-definitie 119
- pascal-definitie 118
- percentiel 20,29
- permeabiliteit 123
- permittiviteit 123
- peta - voorvoegsel 117
- phot-definitie 120
- pico - voorvoegsel 117
- pint-definitie 119
- planck - const. van 123
- plotten meetpunten 46
- plotten met foutmarkering 46
- poise-definitie 120
- Poissonverdeling 21, 87
- pound-definitie 119
- poundforce-definitie 119
- proton - g-factor 123
- proton - massa 123
- Q**
- quart-definitie 119
- R**
- röntgen-definitie 120
- rad-definitie 120
- rang-gebaseerde methoden 42
- rank-based methoden 41
- regressiecoëfficiënt 60
- relatieve fouten 13
- rem-definitie 120
- residuen 70
- rms deviation 19
- robuuste schattingen 40
- S**
- samengestelde fouten 14
- schatting gemiddelde 33
- schatting variantie 33
- schattingen - robuust 40
- seconde - definitie 117
- SI eenheden 9, 117
- siemens-definitie 118
- sievert-definitie 118
- significante afwijking 23,41
- significante cijfers 8
- skewness 19
- skewness-schatting 36
- standaarddeviatie 19

standaarddeviatie-nauwkeurigheid 35,96yotta - voorvoegsel 117

standaarddeviatie-schatting 33

stefan-boltzmann const. 123

stilb-definitie 120

stokes-definitie 120

stone-definitie 119

Student's t-verdeling 34,89

Student's t-verdeling - cumulatief 136

Student's t-verdeling - zakb. 135

sufficient statistics 89

systematische fout 12,83

### T

t-verdeling 34

t-verdeling - cumulatief 136

t-verdeling - zakb. 135

tera - voorvoegsel 117

tesla-definitie 118

toevallige fouten 12

ton-definitie 119

torr-definitie 120

### U

uitschieters 41

### V

variantie 18

variantie-nauwkeurigheid 35,96

variantie-schatting 33,91

verwachte waarde 18

verwachting 18

vrijheidsgraden 34,68,102

### W

waarschijnlijkheidsdichtheid 17

waarschijnlijkheidspapier 29

waarschijnlijkheidsverdeling 18

watt - definitie 118

weber - definitie 118

weegfactoren-ongelijk 97

### Y

yard - definitie 119

yocto - voorvoegsel 117

### Z

zeemijl - definitie 119

zepto - voorvoegsel 117

zetta - voorvoegsel 117

zuivere schatting 101

zwaartepuntsmethode 53