Statistiek voor fysici

Prof. Jorgen D'Hondt
Interuniversitair Instituut voor Hoge Energieën
Vrije Universiteit Brussel

Vakgroep Natuurkunde
Faculteit van de Wetenschappen
Vrije Universiteit Brussel

Inhoud

Voorwoord

Deze inleidende cursus 'Statistiek voor fysici' is nieuwe opleidingsonderdeel dat voor het eerst werd gedoceerd tijdens het academiejaar 2006-2007 met de bedoeling de studenten enkele belangrijke concepten eigen te maken, die nodig zijn om de experimentele gegevens te analyseren bij de proeven behorende bij bijvoorbeeld de cursus 'Meten en experimenteren'. Ook komen sommige concepten terug in de studie van de kwantummechanica en de statistische fysica in de hogere jaren van jullie studies.

Het eerste gebruik van de waarschijnlijkheidsleer door de franse wiskundigen Blaise Pascal (1623-1662) en Pierre de Fermat (1601-1665) dateert van de 17^{de} eeuw en begon met het numeriek begrijpen en voorspellen van eenvoudige kansspelen. Vele anderen hebben deze bevindingen in een wiskundig kader geplaatst zoals we de waarschijnlijkheidsleer of kansrekening tot op heden formuleren en in ons dagelijks leven gebruiken. Welk weer wordt het morgen ? Wat is de kans dat mijn ziekte geneest met deze nieuwe behandeling ? Wie wint er de Tour de France volgend jaar onder voorwaarde dat men geen doping gebruikt ? Wat is de slaagkans in de eerste Bachelor Natuurkunde ?

De tak van de wetenschap die men als Statistiek beschrijft, omvat de studie van methoden en technieken om via experimentele metingen informatie te verzamelen over de waarschijnlijkheid dat een fysische grootheid een zekere waarde aanneemt. De Statistiek gaat uit van een verzameling metingen die de wetten van de waarschijnlijkheidsleer volgen.

In deze cursus stappen we af van de gebruikelijke veronderstelling dat we een meting oneindig veel keer kunnen herhalen, zoals vaak door wiskundigen wordt aangenomen. Realistischer dan deze asymptotische eigenschappen hebben de fysici meestal te maken met een eindige set van experimentele metingen.

De eerste zes hoofdstukken behandelen de theorie van de waarschijnlijkheid die enkele essentiële elementen introduceert om een statistische verwerking van experimentele gegevens mogelijk te maken. Na het inleiden van het klassieke concept waarschijnlijkheid, ligt de nadruk op de beschrijvende numerieke of grafische weergave van experimentele meetresultaten.

De hieropvolgende drie hoofdstukken omschrijven enkele belangrijke technieken en concepten van de statistiek: "Hoe een fysische grootheid gemeten kan worden, hoe het resultaat weergegeven moet worden en hoe men deze resultaten moet interpreteren".

Achteraan in de syllabus kan je de referentie vinden van enkele uitstekenden tekstboeken, dit omdat deze inleidende cursus zeker niet de betrachting heeft volledig te zijn. De studenten die zich willen verdiepen in de leerstof worden ofwel naar deze literatuur verwezen, ofwel naar cursussen in de volgende jaren van hun opleiding tot fysicus.

De elektronische versie van de syllabus is te vinden op de website van het Interuniversi-

tair Instituut voor Hoge Energieën of IIHE, `http://w3.iihe.ac.be`, onder de rubriek Onderwijs (VUB) - Cursussen. Bij mijn naam kan je de cursus downloaden. Je kan mij ook steeds contacteren via email, `jodhondt@vub.ac.be`. De oefeningen worden dit jaar gegeven door drs.Cedric Lemaitre (`celemait@vub.ac.be`). Hij zal jullie leren hoe je de concepten van het hoorcollege in de praktijk kan gebruiken. Dit omhelst zowel rekenoefeningen op papier, als complexe verwerkingen met behulp van het software pakket *Mathematica*. Aarzel niet om regelmatig je onopgeloste vragen te stellen over de leerstof zodat je aan het eind van het jaar niet voor problemen komt te staan.

Hoofdstuk 1

Basisbegrippen van de waarschijnlijkheid

"If any subject might be expected to baffle the mathematicians, it would be chance. In truth, the notion of chance, probability, likelihood, or by whatever name it may be called, is as much of its own nature the object of mathematical reasoning, as force or colour: it contains in itself a distinct application of the notion of relative magnitude: it is 'more' or 'less', and the only difficulty (as in many other cases) lies in the assignment of the test of quantity, 'how much' more or less."

P.-S. le Marquis de Laplace,
'Théorie Analytique des Probabilités', Paris 1820

Het concept van een waarschijnlijkheid of kans is op het eerste zicht een zeer intuïtief begrip. Men kan eenvoudig de kans bepalen om met een muntstuk 'kop' of 'munt' te gooien. Niettegenstaande deze ogenschijnlijke simpliciteit heeft men er verschillende eeuwen over gedaan om een consistent model op te stellen voor de waarschijnlijkheidsleer. Want neem nu eens dat het bovenvermelde muntstuk niet volledig symmetrisch is en bijgevolg 'vervalst' is. De kans om beide uitkomsten te bekomen is bijgevolg ook niet meer symmetrisch en onze eenvoudige definitie van waarschijnlijkheid moeten we herzien.

1.1 Definitie van de waarschijnlijkheid

Hedendaags bestaat er geen unieke definitie van waarschijnlijkheid die algemeen aanvaard wordt. In het algemeen kunnen we de verschillende definities in twee categoriëen onderscheiden : de klassieke- of frequentiebenadering en de meer moderne versie, gebaseerd op de verzamelingsleer. De eerste benadering associeert waarschijnlijkheid met de uitkomst van

herhaalde metingen of experimenten, terwijl de tweede benadering het begrip waarschijnlijkheid associeert met een graad van kennis.

1.1.1 Klassieke definitie

Beschouw een experiment met n verschillende uitkomsten. Indien we in totaal N gelijkaardige experimenten uitvoeren, kunnen we de waarschijnlijkheid of kans $P(X_i)$ ($i \in \{1, ..., n\}$) definiëren om uitkomst X_i te verkrijgen alszijnde de verhouding van het aantal keer we X_i bekomen, namelijk n_{x_i} tot het totaal aantal keer we het experiment herhaald hebben

$$\lim_{N\to\infty} \frac{n_{x_i}}{N} = P(X_i) \tag{1.1}$$

waar een fysicus meestal uitgaat van de limietwaarde, zelfs al is die a priori niet gekend. Voor een wiskundige (alsook in deze cursus) is deze definitie moeilijk toe te passen, omdat hij gebaseerd is op het uitvoeren van oneindig veel experimenten en omdat het gebruik van de waarschijnlijkheidsleer beperkt wordt tot grootheden die meetbaar zijn. Hierbij is het essentieel dat alle n mogelijke uitkomsten een gelijke kans hebben en dus even waarschijnlijk zijn

$$P(X_1) = P(X_2) = ... = P(X_n) \ . \tag{1.2}$$

Als men in een casino een dobbelsteen werpt, gaat men ervan uit dat de zes mogelijke uitkomsten een gelijke kans van voorkomen hebben (of indien men een 'eerlijke' casino binnenwandelt). Deze definitie, ingevoerd door Laplace, kunnen we relateren aan het begrip *frequentie* van voorkomen. De frequentie of waarschijnlijkheid om een '2' te werpen is 1/6.

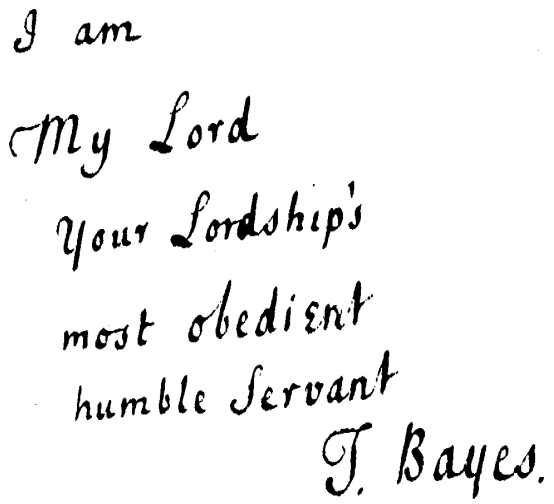
Let hierbij dat de definitie van de waarschijnlijkheid reeds gebruik maakt van de term 'waarschijnlijkheid' (gelijke kans van voorkomen voor de 6 zijden van de dobbelsteen). We moeten deze definitie dus beschouwen als een methode om de waarschijnlijkheid te bepalen en niet als een volledige definitie van het begrip.

1.1.2 Moderne definitie

Een meer volledige definitie van waarschijnlijkheid werd geïntroduceerd door Thomas Bayes (1702-1761). Indien we niet met absolute zekerheid kunnen verklaren of een gebeurtenis juist of vals is, moeten we zeggen dat deze 'meer' of 'minder' waarschijnlijk is. Op die manier kunnen we verschillende graden van waarschijnlijkheid onderscheiden indien we (eventueel mits voorkennis) kunnen inschatten of de gebeurtenis meer juist of meer fout is. Wetende dat de frequentiewaarschijnlijkheid moet voldoen aan enkele basisregels die men eenvoudig rechtstreeks kan bewijzen uit deze frequentiedefinitie, willen we ook deze 'moderne' definitie van waarschijnlijkheid dezelfde regels opleggen. Deze regels noemen we dan axioma's, welke we opstellen binnen het kader van de verzamelingsleer.

Een modernere definitie van waarschijnlijkheid is gebaseerd op de verzamelingsleer. Definieer de steekproefruimte Ω als de verzameling van alle mogelijke gebeurtenissen of deelruimten X_i waar elk van deze gebeurtenissen $i \in \{1, ..., n\}$ exclusief is (als er één voorkomt komen de andere niet voor of $\forall i, j \in \{1, ..., n\}$ en $i \neq j$ geldt $X_i \cap X_j = \emptyset$)

REV. T. BAYES



$$\Omega \equiv \left(\bigcup_{i=1}^{n} X_i \right) \quad . \tag{1.3}$$

De waarschijnlijkheid $P(X_i)$ dat gebeurtenis X_i voorkomt is dan een reëel getal dat voldoet aan volgende axioma's opgesteld door Andrei Nikolaevich Kolmogorov (1903-1987)

- $0 \leq P(X_i) \leq 1$
- P(Ω) = 1
- $\forall$ i,j$\in \{1, ..., n\}$ en $i \neq j : P(X_i \cup X_j) = P(X_i) + P(X_j)$

Het eerste axioma legt beperkingen op de waarschijnlijkheid van iedere deelverzameling X_i. Deze moet een reëel getal zijn tussen 0 en 1. Het tweede vertelt ons dat de totale kans altijd gelijk moet zijn aan 1. Het derde axioma geldt voor exclusieve of niet overlappende deelverzamelingen en kan men veralgemenen naar

$$P(\bigcup_{i=1}^{n} X_i) = \sum_{i=1}^{n} P(X_i) \quad . \tag{1.4}$$

Deze moderne definitie van waarschijnlijkheid komt overeen met de klassieke frequentie definitie indien men de relatieve frequentie van voorkomen bepaalt via een experiment dat men oneindig keer herhaalt.

De moderne definitie heeft het voordeel dat men kan spreken over de waarschijnlijkheid van de echte waarde van een fysische grootheid of over de waarschijnlijkheid dat een theorie juist of fout is. De frequentie definitie kan enkel iets beweren over de waarschijnlijkheid van de uitkomst van een experiment, welke gebeurtenis meer of minder frequent voorkomt.

1.2 Eigenschappen van de waarschijnlijkheid

Via de frequentie interpretatie van de waarschijnlijkheid kunnen we enkele eigenschappen opstellen met betrekking tot het rekenen met waarschijnlijkheden. Beschouw twee verschillende verschijnselen A en B, die niet exclusief zijn (ze kunnen dus samen optreden). Indien men n waarnemingen maakt, komt men in vier verschillende situaties uit :

- verschijnsel A treedt op en verschijnsel B treedt niet op (n_1 maal)
- verschijnsel B treedt op en verschijnsel A treedt niet op (n_2 maal)
- beide verschijnselen A en B treden op (n_3 maal)
- geen van beide verschijnselen A of B treden op (n_4 maal)

De getallen n_i met $i \in \{1, 2, 3, 4\}$ geven de frequentie van voorkomen weer voor deze vier gebeurtenissen, met de relatie

$$\sum_{i=1}^{4} n_i = n \quad . \tag{1.5}$$

Voor de relatieve frequentie van het voorkomen van het verschijnsel A bekomen we

$$f(A) = \frac{n_1 + n_3}{n} \, , \tag{1.6}$$

terwijl we voor het verschijnsel B de volgende uitdrukking bekomen

$$f(B) = \frac{n_2 + n_3}{n} \quad . \tag{1.7}$$

De relatieve frequentie voor het optreden van hetzij A, hetzij B, hetzij beide, symbolisch voorgesteld door (A+B), is

$$f(A + B) = \frac{n_1 + n_2 + n_3}{n} \quad . \tag{1.8}$$

De relatieve frequentie voor het optreden van A én B, hetgeen symbolisch geschreven wordt als AB, is

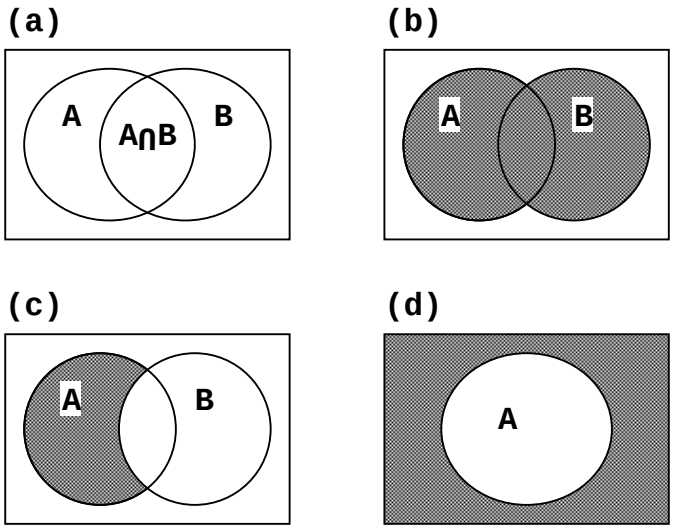
$$f(AB) = \frac{n_3}{n} \tag{1.9}$$

Bijgevolg voldoen de relatieve frequenties aan de volgende relatie

$$f(A + B) = f(A) + f(B) - f(AB) \quad . \tag{1.10}$$

De waarschijnlijkheid dat ten minste één der twee verschijnselen A en B optreedt is gelijk aan de som van beide individuele waarschijnlijkheden min de waarschijnlijkheid dat beide verschijnselen gelijktijdig optreden. Indien A en B exclusieve verschijnselen zijn, vereenvoudigt de relatie zich tot

$$f(A + B) = f(A) + f(B) \quad . \tag{1.11}$$



Figuur 1.1: *Basis operatoren met verzamelingen : (a) doorsnede $A \cap B$, (b) unie $A \cup B$, (c) verschil $A - B$ en (d) complement $\overline{A}$ (de donkere delen duiden deze begrippen aan).*

Deze verschillende categorieën van observaties kunnen we ook voorstellen met verzamelingen in Figuur 1.1.

De doorsnede van twee deelverzamelingen A en B van de steekproefruimte Ω schrijven we als

$$A \cap B = \{x | x \in A \text{ en } x \in B\} \tag{1.12}$$

en de unie als

$$A \cup B = \{x | x \in A \text{ of } x \in B\} \tag{1.13}$$

Het verschil kunnen we schrijven als

$$A - B = \{x | x \in A \text{ en } x \notin B\} \tag{1.14}$$

en het complement als

$$\overline{A} = \{x | x \in \Omega \text{ en } x \notin A\} \ . \tag{1.15}$$

Met deze definities van verschijnselen alszijnde deelverzameling van een steekproefruimte Ω en met de axioma's van de waarschijnlijkheid, kunnen we overgaan van relatieve frequenties $f(X)$ naar moderne waarschijnlijkheden $P(X)$. Ook voor de moderne definitie van waarschijnlijkheid willen we dergelijke regels opleggen

$$P(A + B) = P(A) + P(B) - P(AB) \ . \tag{1.16}$$

De optellingswet 1.16 kunnen we uitbreiden voor meerdere deelverzamelingen A_i van de steekproefruimte Ω. Door steeds opnieuw dezelfde relatie te gebruiken kunnen we volgende algemene relatie bewijzen

$$P(A_1 + A_2 + ... + A_n) = S_1 - S_2 + S_3 - ... - (-1)^n S_n \tag{1.17}$$

met volgende definities

$$S_1 = \sum_{i=1}^{n} P(A_i) \tag{1.18}$$

$$S_2 = \sum_{i=1}^{n} \sum_{j=1}^{i-1} P(A_i A_j) \tag{1.19}$$

$$S_3 = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} P(A_i A_j A_k) \tag{1.20}$$

$$... \tag{1.21}$$

$$S_n = P(A_1 A_2 A_3 ... A_n) \tag{1.22}$$

waarbij de notatie $A_1 + A_2 + ... + A_n$ betekent dat ten minste één der n verschijnselen $A_1, A_2, ..., A_n$ optreedt, terwijl het symbool $A_1 A_2 ... A_n$ betekent dat de n verschijnselen gelijktijdig optreden.

Met het symbool $\overline{\mathrm{A}}$ bedoelen we alle mogelijke gebeurtenissen in de steekproefruimte Ω die niet in de verzameling van gebeurtenissen A zitten. Men heeft bijgevolg twee eenvoudige relaties

$$P(A) + P(\overline{A}) = P(\Omega) = 1 \tag{1.23}$$

en

$$P(A\overline{A}) = 0 \tag{1.24}$$

en ook de volgende relatie kan men bewijzen

$$P(A + B) = 1 - P(\overline{AB}) \ . \tag{1.25}$$

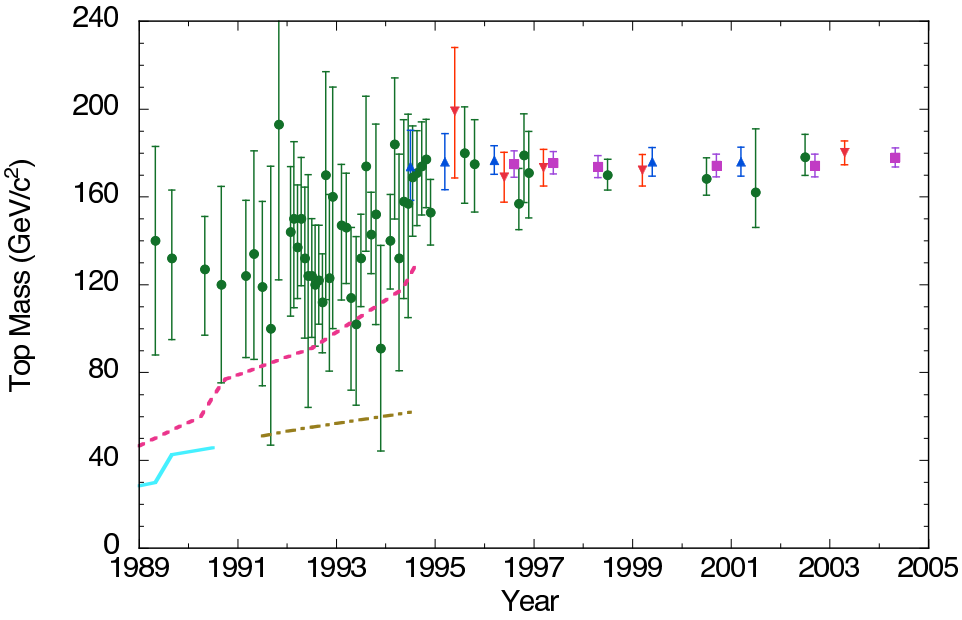
Bewijs relatie 1.25: Stel A = AB + $A\overline{B}$ welke de som is van twee exclusieve verzamelingen, dan geldt via de optellingswet 1.16 dat P(A) = P(AB) + P($A\overline{B}$). Bijgevolg is P(A+B) = P(B) + P($A\overline{B}$) = 1 - P($\overline{B}$) + P($A\overline{B}$) en is $\overline{B}$ = $\overline{B}A$ + $\overline{AB}$, dus P($\overline{B}$) = P($\overline{B}A$) + P($\overline{BA}$) en geldt door beide relaties samen te nemen de relatie P(A+B) = 1 - P($\overline{B}A$) - P($\overline{BA}$) + P($A\overline{B}$) = 1 - P($\overline{AB}$). QED [1]

De relatie tussen de waarschijnlijkheid van een gebeurtenis of een verzameling van gebeurtenissen en de waarschijnlijkheid van het complement is nuttig bij het bepalen van kansen. De waarschijnlijkheid P om minstens één 6 te gooien bij vier worpen met eenzelfde dobbelsteen wordt berekend via het complementair verschijnsel: de gebeurtenis waarbij men géén 6 gooit in één der vier worpen. We bekomen bijgevolg de waarschijnlijkheid als P = 1 - $(\frac{5}{6})^4$ = 0.518. Een ander voorbeeld van het gebruik van relatie 1.23 is het berekenen van de waarschijnlijkheid P' om minstens één dubbele 6 te gooien bij 24 worpen met twee dobbelstenen, namelijk P' = 1 - $(\frac{35}{36})^{24}$ = 0.491.

[1] QED = Quod erat demonstrandum.

1.3 Conditionele waarschijnlijkheid

De moderne definitie van waarschijnlijkheid is een subjectief begrip. Als men iemand vraagt naar hoeveel geloof hij of zij hecht aan het feit dat of waarschijnlijkheid hij of zij geeft aan het feit dat morgen de zon zal schijnen boven Brussel, verkrijgt men een antwoord afhankelijk van de persoon aan wie de vraag gesteld is. Zo zal een Spaanse toerist in België een optimistische kans van 0.9 geven voor deze gebeurtenis, terwijl de Belg zelf een realistische kans van 0.6 zal toekennen aan de gebeurtenis. De waarschijnlijkheid P(X) voor de gebeurtenis X is bijgevolg geen intrinsieke eigenschap van de gebeurtenis, maar is afhankelijk van de informatie waarover de persoon beschikt om P(X) te bepalen. We spreken bijgevolg van een subjectieve definitie. Neem de deelverzamelingen van de steekproefruimte van gebeurtenissen als hypotheses die juist of fout kunnen zijn. Voor de 'frequentisten' ligt de meting of observatie altijd binnen of buiten een bepaalde deelverzameling. Volgens de subjectieve waarschijnlijkheid geeft men een graad van 'geloof' aan het feit dat de meting erbinnen of erbuiten ligt.



Figuur 1.2: *Evolutie van de kennis in verband met de massa van de top-quark binnen het Standaard Model van de elementaire deeltjes. De grafiek geeft de meest waarschijnlijke waarde van de top-quark massa weer met zijn onzekerheid (zie verder in de cursus), dit als functie van het jaar waarin men deze waarschijnlijkheid heeft geëvalueerd.*

Als de voorkennis verandert, zal bijgevolg ook de evaluatie van de waarschijnlijkheid van het voorkomen van een gebeurtenis veranderen. Dit kunnen we duidelijk maken aan de hand van de metingen van de top-quark massa die de laatste decennia werden uitgevoerd. De gebeurtenissen waarvoor men een waarschijnlijkheid moet bepalen zijn: 'de top-quark massa is gelijk aan xx' en dit voor elke reële waarde van xx. De top-quark is een heel zwaar elementair deeltje dat enkel kan geproduceerd worden door het samenbrengen van veel energie (zie cursus 'Elementaire Deeltjes Fysica' in het derde Bachelor jaar). Tot voor kort was

geen enkel laboratorium in de wereld erin geslaagd om dit deeltje kortstondig aan te maken en bijgevolg een meting mogelijk te maken. Voor zijn ontdekking in 1995 moest men er dus vanuit gaan dat het deeltje bestond via zijn voorspelling in de theorie van het Standaard Model. Door de indirecte effecten van het deeltje te bestuderen kon men afleiden dat het deeltje een massa moest hebben die groter is dan ongeveer 50 GeV/c^2 [2] (de stippellijn op Figuur 1.2) en dit was een hypothese waar men veel geloof aan hechtte. De waarschijnlijkheid van de gebeurtenis 'de top-quark massa is groter dan 50 GeV/c^2' is veel groter dan de gebeurtenis 'de top-quark massa is kleiner dan 50 GeV/c^2'. Door de metingen van deze indirecte effecten dacht men met een zekere hoeveelheid geloof dat de top-quark massa ongeveer 120 GeV/c^2 was. Maar men hechtte bijna evenveel geloof (juist iets minder) aan een waarde van 150 GeV/c^2 of 90 GeV/c^2. Met hun voorkennis in acht genomen, konden de wetenschappers niet exact bepalen welke waarde van de top-quark massa het meest waarschijnlijk was. In 1995 hebben onderzoekers, verbonden aan Fermilab nabij Chicago (V.S.), voor het eerst enkele van deze top-quarks rechtstreeks kunnen maken en bestuderen. Ze hadden hun kennis in verband met de top-quark massa dus significant uitgebreid, met als gevolg dat ze nu met een grotere nauwkeurigheid kunnen bepalen welke waarde van de top-quark massa het meest waarschijnlijk is. Hun, met de tijd evoluerende voorkennis, liet toe om de evaluatie van de waarschijnlijkheid van de juiste waarde van de top-quark massa te herzien. Nu hecht men veel geloof aan de gebeurtenis 'de top-quark massa is 178 GeV/c^2' terwijl men veel minder geloof hecht aan de gebeurtenissen 'de top-quark massa is 170 GeV/c^2' of 'de top-quark massa is 186 GeV/c^2'.

Deze waarneming leidt ons naar het begrip van conditionele waarschijnlijkheid P(A|B), de waarschijnlijkheid dat gebeurtenis A waar is indien men met zekerheid weet dat gebeurtenis B waar is. Bijvoorbeeld de kans om 'kop' te bekomen indien gegeven is dat we met een 'eerlijke munt' gooien, kunnen we noteren als P('kop'|'eerlijke munt') = 0.5.

Stel dat de steekproefruimte bestaat uit $\Omega = \{x_1, x_2, ..., x_n\}$. We kunnen dus met elk element van de steekproefruimte een frequentie van voorkomen $f(x_i)$ (i$\in \{1, ..., n\}$) associëren. Stel nu dat gebeurtenis Y reeds waar is, met Y een deelverzameling is van Ω ($Y \subset \Omega$). Met deze voorkennis moeten we de frequenties $f(x_i)$ herzien en de conditionele frequenties $f(x_i|Y)$ bepalen. Uiteraard willen we bekomen dat indien $x_j \notin Y$, dat $f(x_j|Y) = 0$. Ook willen we dat de relatieve frequenties van de elementen in de steekproefruimte Ω die ook in de deelverzameling Y zitten, gelijk blijft. We hebben namelijk geen informatie die kan differentiëren tussen de elementen in Y. Bijgevolg bekomen we de relatie

$$f(x_i|Y) = C \cdot f(x_i) \quad \forall x_i \in Y \tag{1.26}$$

met C een positieve constante ($C \geq 1$). Ook willen we dat de som van alle frequenties 1 is, of

$$\sum_{i \in Y} f(x_i|Y) = C \cdot \sum_{i \in Y} f(x_i) = 1 \tag{1.27}$$

of

[2] De eenheid van Electron Volt (eV) is de energie die een elektron heeft als het in een lineair potentiaalverschil van 1 Volt versneld wordt. Giga Electron Volt is 10^9 keer een Electron Volt.

$$C = \frac{1}{\sum_{i \in Y} f(x_i)} = \frac{1}{P(Y)} \tag{1.28}$$

waar $P(Y)$ de kans is dat gebeurtenis Y voorkomt ($P(Y) > 0$). Dit geeft ons volgende uitdrukking voor de conditionele frequenties

$$f(x_i|Y) = \frac{f(x_i)}{P(Y)} \quad \forall x_i \in Y \ . \tag{1.29}$$

Om deze logische uitdrukking binnen het kader van de frequentie of klassieke definitie van de waarschijnlijkheid over te nemen naar de moderne definitie, moeten we de relatie postuleren als een axioma. De waarschijnlijkheid dat de twee verschijnselen X en Y optreden, is gelijk aan het product van de waarschijnlijkheid van het ene verschijnsel met de conditionele waarschijnlijkheid van het andere verschijnsel

$$P(X|Y) = \sum_{i \in (X \cap Y)} f(x_i|Y) = \sum_{i \in (X \cap Y)} \frac{f(x_i)}{P(Y)} = \frac{P(X \cap Y)}{P(Y)} = \frac{P(XY)}{P(Y)} \ . \tag{1.30}$$

Voor meer dan twee gebeurtenissen geldt:

$$P(X_1 X_2 ... X_k) = P(X_1) \cdot P(X_2|X_1) \cdot ... \cdot P(X_k|X_1 X_2 ... X_{k-1}) \ . \tag{1.31}$$

Gebeurtenissen $X \subset \Omega$ ($P(X) > 0$) en $Y \subset \Omega$ ($P(Y) > 0$) zijn onafhankelijk als en slechts als

$$P(XY) = P(X) \cdot P(Y) \ . \tag{1.32}$$

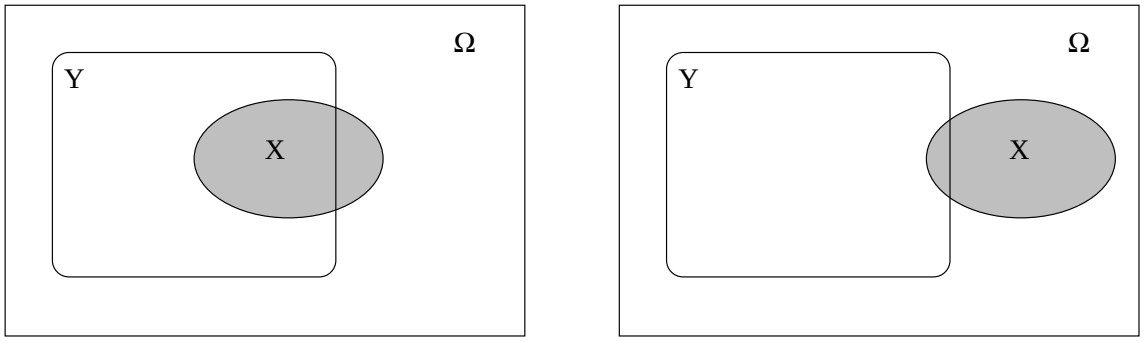
Met andere woorden heeft de waarschijnlijkheid van het voorkomen van gebeurtenis X geen invloed op de waarschijnlijkheid van het voorkomen van gebeurtenis Y en omgekeerd.

Bewijs relatie 1.32: Veronderstel dat gebeurtenissen X en Y inderdaad onafhankelijk zijn. Dan geldt voor de conditionele waarschijnlijkheid P(X|Y) = P(X) en dus P(XY) = P(X ∩ Y) = P(X|Y) · P(Y) = P(X) · P(Y). Omgekeerd kunnen we veronderstellen dat de relatie P(XY) = P(X) · P(Y) waar is. Bijgevolg is P(X|Y) = P(X ∩ Y) / P(Y) = P(X) en P(Y|X) = P(Y ∩ X) / P(X) = P(Y) en zijn beide gebeurtenissen onafhankelijk. QED

Dit kunnen we ook veralgemenen. De gebeurtenissen $\{X_1, X_2, ... , X_k\}$ zijn ieder onafhankelijk ten opzichte van elke combinatie van de anderen als en slechts als voor elke eindige deelverzameling van $r \leq k$ gebeurtenissen geldt

$$P\left(\bigcap_{i=1}^{r} X_i\right) = \prod_{i=1}^{r} P(X_i) \ . \tag{1.33}$$

Indien echter $P(X|Y) \neq P(X)$ zijn de gebeurtenissen niet onafhankelijk en bijgevolg gecorreleerd. We spreken van een positieve correlatie tussen gebeurtenis X en Y indien $P(X|Y) > P(X)$, en een negatieve correlatie indien $P(X|Y) < P(X)$. Bij een positieve correlatie zal het voorkomen van gebeurtenis Y de waarschijnlijkheid verhogen voor het



positieve correlatie tussen X en Y negatieve correlatie tussen X en Y

Figuur 1.3: *Illustratie van een positieve en negatieve correlatie tussen gebeurtenissen.*

voorkomen van gebeurtenis X. Dit kunnen we eenvoudig duidelijk maken aan de hand van Figuur 1.3.

Indien de relatieve waarschijnlijkheid van de gebeurtenis X ten opzichte van $\overline{X}$ groter is binnen de deelverzameling Y vergeleken met dezelfde verhouding binnen de globale steekproefruimte Ω, spreken we van een positieve correlatie. Een negatieve correlatie verkrijgen we in het omgekeerde geval.

1.4 De regel van Bayes

Beschouw een verzameling van exclusieve gebeurtenissen of hypothesen $\{X_i \mid i \in \{1, ..., n\}\}$ die samen de steekproefruimte vormen zodat $\Omega = \bigcup_{i=1}^{n} X_i$. Er geldt dus dat $P(X_i X_j) = 0$ indien $i \neq j$. Dit kan bijvoorbeeld zijn: 'het is 13 uur', 'het is 17 uur' of 'het is 2 uur'. Voor we een experiment beginnen of nieuwe informatie ontvangen, kunnen we onze kennis over de waarschijnlijkheden van de verschillende gebeurtenissen of hypothesen samenvatten als $\{P(X_i) \mid i \in \{1, ..., n\}\}$. We kunnen deze waarschijnlijkheden beschouwen als '*a priori*' kennis (voorkennis). Bijvoorbeeld met de voorkennis dat we ons in België bevinden en dat de zon schijnt, weten we dat de waarschijnlijkheid voor de gebeurtenis 'het is 2 uur' nul is. Het kan ook dat we helemaal geen voorkennis hebben, wat resulteert in gelijke kansen voor elk van de gebeurtenissen X_i, of $P(X_i) = 1/n$. Hierna creëren we nieuwe informatie door het uitvoeren van een experiment waarin we gebeurtenis Y waarnemen. Het verschijnsel Y kan n verschillende oorzaken hebben, namelijk een van de hypothesen in de verzameling $\{X_i \mid i \in \{1, ..., n\}\}$. Wetende dat het verschijnsel Y zich heeft voorgedaan, willen we bepalen wat de waarschijnlijkheid was dat de oorzaak X_i hiervoor de aanleiding was. Deze oorzaken X_i kunnen we bijgevolg beschouwen als hypothesen die het verschijnsel of de gebeurtenis Y hebben bewerkstelligd. Het experiment kan bijvoorbeeld resulteren in de observatie dat de zon ongeveer in het Zuiden staat. We willen nu de extra informatie van het observeren van verschijnsel Y gebruiken om onze kennis over de oorzaken X_i te vergroten. Met andere woorden willen we nagaan welke de conditionele waarschijnlijkheid $P(X_i|Y)$ is. Om deze conditionele waarschijnlijkheden te vinden, kunnen we de volgende relaties gebruiken

$$P(X_i|Y) = \frac{P(X_i \cap Y)}{P(Y)} \tag{1.34}$$

of omgekeerd

$$P(Y|X_i) = \frac{P(Y \cap X_i)}{P(X_i)} \quad . \tag{1.35}$$

Omdat slechts één van de hypothesen X_i juist kan zijn, kunnen we schrijven dat

$$P(Y) = P(X_1 \cap Y) + P(X_2 \cap Y) + ... + P(X_n \cap Y) = \sum_{i=1}^{n} P(X_i \cap Y) \quad . \tag{1.36}$$

Door steeds de relatie 1.35 te gebruiken in uitdrukking 1.36 bekomen we:

$$P(Y) = \sum_{i=1}^{n} P(X_i) \cdot P(Y|X_i) \tag{1.37}$$

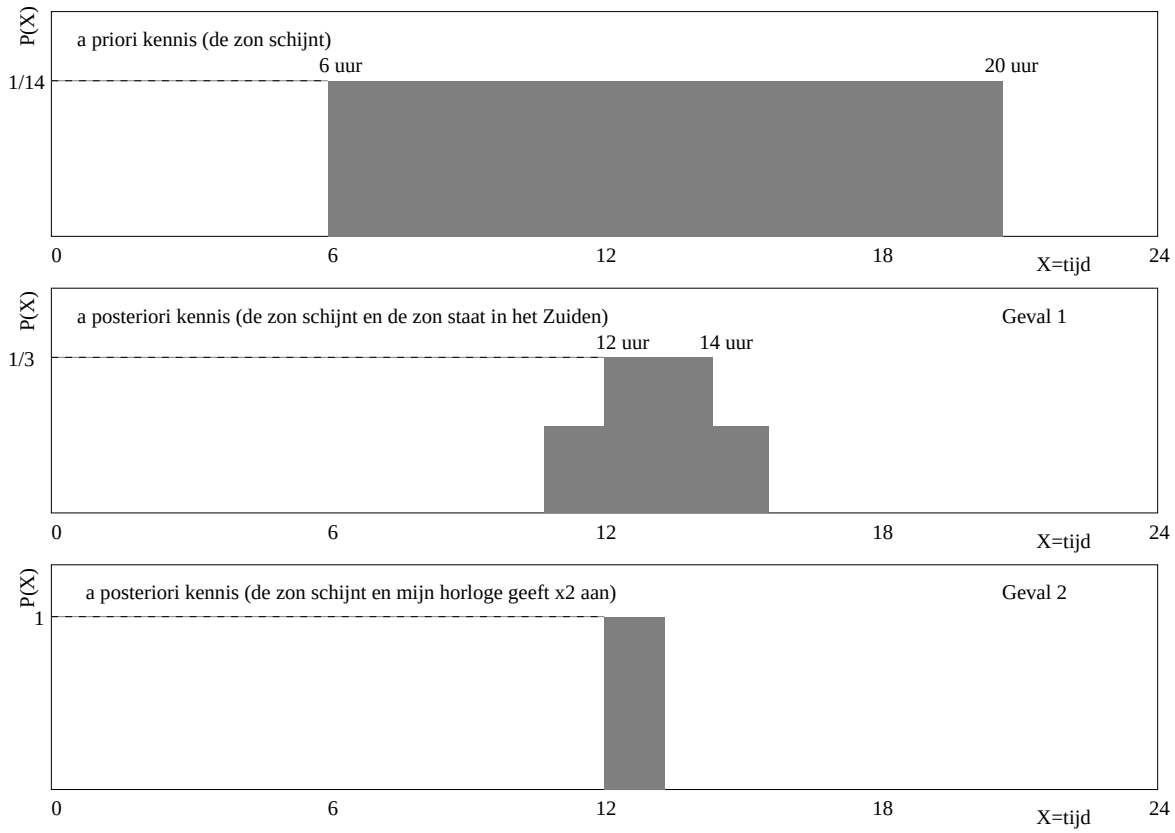
en via relatie 1.34 bekomen we:

$$P(X_i|Y) = \frac{P(X_i) \cdot P(Y|X_i)}{\sum_{i=1}^{n} P(X_i) \cdot P(Y|X_i)} \quad . \tag{1.38}$$

Vergelijking 1.38 noemen we de Regel van Bayes. Deze regel laat toe om de waarschijnlijkheden $P(X_i|Y)$ na het experiment te bepalen, we spreken van '*a posteriori*' waarschijnlijkheden. Wetende dat de zon ongeveer in het Zuiden staat zullen de *a posteriori* waarschijnlijkheden voor de hypothese 'het is 13 uur' veel hoger liggen dan de *a posteriori* waarschijnlijkheden voor de hypothese 'het is 17 uur'.

Een eenvoudige toepassing van de Regel van Bayes is geïllustreerd in Figuur 1.4. De set van 24 gebeurtenissen bestaat uit 'het is X_i uur' waar X_i loopt van 0 tot en met 23 (laat ons eenvoudigweg de minuten vergeten). We voeren een experiment of meting uit met de voorkennis dat de zon schijnt (in België !). In de bovenste grafiek vinden we dus een samenvatting van de waarschijnlijkheden $P(X_i)$ rekening houdend met deze voorkennis. Dit noemen we soms ook de '*prior*' verdeling. In een eerste geval bepaalt ons experiment de stand van de zon met een zekere nauwkeurigheid. Het experiment leert ons dat de zon ongeveer in het Zuiden staat (gebeurtenis Y). Nu weten we dat indien de zon in het Zuiden staat, het 13 uur is. Hiermee kunnen we *a posteriori* waarschijnlijkheden bepalen $P(X_i|Y)$. Omdat we slechts een ruwe meting gedaan hebben van de stand van de zon, kunnen we niet eenduidig zeggen dat het 13 uur is. We moeten dus afnemende waarschijnlijkheden toekennen aan de uren rond 13 uur. Dit wordt geïllustreerd met de tweede grafiek:

- P('11 uur' | 'zon in het Zuiden') = 1/6
- P('12 uur' | 'zon in het Zuiden') = 1/3
- P('13 uur' | 'zon in het Zuiden') = 1/3
- P('14 uur' | 'zon in het Zuiden') = 1/6



Figuur 1.4: *Illustratie van de Regel van Bayes.*

In een tweede geval bepaalt ons experiment het uur aan de hand van mijn digitale horloge. Helaas werkt het eerste digit niet meer. Het tweede digit geeft echter een 2 aan. Rekening houdend met de *a priori* kennis kunnen we de volgende *a posteriori* waarschijnlijkheden opstellen:

- P('12 uur' | 'tweede digit horloge is 2') = 1
- P('niet 12 uur' | 'tweede digit horloge is 2') = 0

Deze waarschijnlijkheden worden weergegeven in de derde grafiek van de figuur. Door middel van de *a priori* kennis kunnen we dus de hypothese uitsluiten 'het is 22 uur'.

In de Regel van Bayes (uitdrukking 1.38) vinden we dat de noemer niets anders is dan een normalisatie factor die gelijk is voor elke *a posteriori* waarschijnlijkheid $P(X_i|Y)$. Indien we die gelijkstellen aan 1 veranderen we niets aan de relatieve verhoudingen van de *a posteriori* waarschijnlijkheden. We kunnen dus vereenvoudigd schrijven:

$$P(X_i|Y) \propto P(Y|X_i) \cdot P(X_i) \tag{1.39}$$

om nadien alle *a posteriori* waarschijnlijkheden te vermenigvuldigen zodat de totale kans $\sum_{i=1}^{n} P(X_i|Y)$ terug 1 wordt.

Wanneer we als fysicus een experiment uitvoeren, willen we in de meeste gevallen onafhankelijk zijn van de zogenaamde *a priori* kennis. Die zou eventueel verkeerd kunnen zijn of misschien willen we die wel verifiëren. Dit is heel eenvoudig mogelijk door alle *a priori* waarschijnlijkheden voor de oorzaken of hypothesen gelijk te stellen, $P(X_i) = P(X_j)$ $(i, j \in \{1, ..., n\})$. We spreken dan over een uniforme *prior* of voorkennisfunctie. In het bovenstaand voorbeeld was de *prior* niet uniform, daar de waarschijnlijkheden voor alle uren vóór 6 uur en ná 20 uur nul waren, terwijl alle uren tussen zonsopgang en zonsondergang een waarschijnlijkheid van 1/14 hadden.

Deze Bayesiaanse manier om waarschijnlijkheden toe te kennen aan verschillende oorzaken is zeer nuttig en is in de praktijk sterk verbonden met het concept van een *likelihood* of een kansfunctie. Deze likelihood heeft een aantal eigenschappen die zeer handig (en misschien iets complexer) zijn bij het analyseren van experimentele gegevens. In de cursus Statistiek in de volgende studiejaren gaan jullie dieper in op de betekenis van dergelijke *likelihood*.

Hoofdstuk 2

Rekenen met waarschijnlijkheden

"No one has caused me any difficulty in regard to the above, but they have told me that they did not do so for the reason that everyone is accustomed to this method today. "

B. Pascal,
Brief van Blaise Pascal aan Pierre de Fermat, 29 juli 1654

De eerste historische toepassing van de regels van de kansrekening vinden we terug in kansspelen. Vroege wiskundigen zoals Blaise Pascal (1623-1662) en Pierre de Fermat (1601-1665) ontworpen verschillende rekenregels om de frequentie te bepalen van de uitkomsten van deze kansspelen. In dit hoofdstuk gaan we dus eenvoudigweg leren tellen, een proces dat niet steeds even gemakkelijk zal blijken !!

2.1 Som- en productregels

De studie van discrete mathematische telproblemen begint bij twee eenvoudige, maar essentiële regels: de som- en productregels. Veel complexe probleemstellingen kunnen namelijk opgesplitst worden in deelproblemen die steeds neerkomen op deze twee basisregels.

- **Somregel**: Als een eerste gebeurtenis op m manieren kan plaatsgrijpen en een tweede gebeurtenis op n manieren, en beide gebeurtenissen kunnen niet tegelijk plaatsgrijpen (exclusieve gebeurtenissen) dan kunnen we één van beide taken op $m + n$ manieren uitvoeren.

- **Productregel**: Als we in een gebeurtenis twee opeenvolgende onderdelen kunnen onderscheiden en als er m manieren zijn voor het eerste onderdeel en n manieren voor het tweede, dan kan de totale gebeurtenis gebeuren op $m \cdot n$ manieren.

Indien er in de algemene bibliotheek aan de Vrije Universteit Brussel 342 boeken aanwezig zijn over het onderwerp Statistiek en er in de bibliotheek van het departement Natuurkunde

nog eens 107 staan, kunnen jullie in totaal kiezen uit 449 boeken om je te verdiepen in het onderwerp Statistiek. Dit in de veronderstelling dat er geen boeken twee keer voorkomen. Eenzelfde dobbelsteen kan twee keer geworpen worden, waarneming X en waarneming Y. De eerste keer zijn er 6 mogelijke uitkomsten voor X, de tweede keer zijn er ook 6 mogelijke uitkomsten voor Y. Bijgevolg zijn er 36 uitkomsten voor de combinatie (X,Y), waarvan er sommige gelijk zijn. Indien we twee dobbelstenen met verschillende kleuren hebben, zijn alle 36 uitkomsten verschillend.

In de volgende twee onderdelen gaan we leren hoe we een complex probleem moeten opsplitsen in deelproblemen. Nadien moeten we de oplossing van deze deelproblemen samennemen om een antwoord te formuleren op het initieel complex probleem.

2.2 Permutaties

Om de volgende probleemstelling in te leiden, kunnen we deze vraag bekijken: Hoeveel personen hebben we minimum nodig in deze klas om een kans te hebben groter dan 0.5 dat twee personen dezelfde verjaardag hebben? Indien we veronderstellen dat er 365 dagen in een jaar zijn, zullen sommige misschien denken dat het antwoord simpelweg de helft van 365 of ongeveer 183 is. Met zoveel mensen zullen ze bijna zeker twee personen vinden die dezelfde verjaardag hebben. Om het correcte antwoord te vinden moeten we de personen rangschikken van 1 tot r, waar r het aantal personen is in de klas. De eerste persoon heeft 365 mogelijke verjaardagen, zo ook alle andere. Dit geeft in totaal 365^r mogelijke sequenties van verjaardagen. Uit deze moeten we diegene vinden waarin geen gelijke verjaardagen voorkomen. Bijgevolg heeft de eerste nog steeds 365 mogelijkheden, de tweede slecht 364, de derde 363, enzovoort. Persoon r heeft bijgevolg nog 365-r+1 mogelijkheden. Het totaal aantal sequenties zonder gelijke verjaardagen is dus

$$365 \cdot 364 \cdot 363 \cdot ... \cdot (365 - r + 1) \tag{2.1}$$

en bijgevolg is de waarschijnlijkheid om dergelijke sequentie aan te treffen in de klas gelijk aan

$$P_r = \frac{365 \cdot 364 \cdot 363 \cdot ... \cdot (365 - r + 1)}{365^r} \quad . \tag{2.2}$$

Je kan eenvoudigweg een kort computerprogramma schrijven om deze vergelijking op te lossen naar r zodat P_r kleiner is dan 0.5. De kleinste waarde van r waarvoor $P_r < 0.5$ geeft het minimum aantal personen die in de klas moeten zitten om een waarschijnlijkheid te hebben groter dan 0.5 dat twee personen op eenzelfde dag verjaren. Het antwoord is 23.

Dit probleem motiveert de volgende definitie van het begrip *permutatie*.

Definitie: Beschouw een verzameling van n objecten, dan is elke lineaire rangschikking van deze objecten een permutatie van deze verzameling.

We kunnen deze definitie ook herformuleren: Is A een eindige verzameling, dan is een permutatie van A een één-bij-één afbeelding op A zelf. De voorwaarde dat een permutatie een één-bij-één afbeelding is, wijst erop dat geen twee elementen van de verzameling A door de afbeelding op éénzelfde element van de verzameling A worden afgebeeld. Een voorbeeld

van een permutatie van de verzameling $\{a, b, c, d\}$ is $\{b, d, a, c\}$, maar niet $\{b, b, d, c\}$. Alle mogelijke permutaties van de verzameling $\{a, b, c\}$ zijn:

$$\{a, b, c\}, \{b, c, a\}, \{c, a, b\}, \{a, c, b\}, \{b, a, c\}, \{c, b, a\} \ . \tag{2.3}$$

Het is eenvoudig om het aantal mogelijke permutaties van een verzameling van n objecten te tellen. Voor het eerste object in de rangschikking kunnen we kiezen uit de n objecten van de verzameling, voor het tweede uit $n-1$, voor het derde uit $n-2$, ... , voor het laatste hebben we geen keuze meer. Bijgevolg kunnen we stellen dat het aantal mogelijke permutaties gelijk is aan

$$n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 3 \cdot 2 \cdot 1 \ . \tag{2.4}$$

Dit leidt tot de invoering van een nieuw concept, namelijk *n faculteit*.

Definitie: Voor elke positief geheel getal n, wordt n faculteit of n! gedefinieerd als

$$n! = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 3 \cdot 2 \cdot 1 \tag{2.5}$$

met de beginvoorwaarde dat 0! = 1.

Let erop dat dit een groot getal kan worden reeds voor kleine waarden van n, bijvoorbeeld 9!=362880. Om dit rekenprobleem te verhelpen, kunnen we gebruik maken van de benaderingsformule van Stirling [1]: n! $\simeq n^n e^{-n}\sqrt{2\pi n}$. Deze formule geldt asymptotisch voor $n \to \infty$.

Met behulp van de definitie van $n!$ kunnen we op een compacte wijze enkele algemene uitdrukkingen bepalen in verband met kansspelen.

Beschouw n verschillende objecten $\{x_1, x_2, ..., x_n\}$ en een positief geheel getal r met $1 \leq r \leq n$. Via de productregel geldt dat het aantal permutaties waarmee we r objecten van de totale n kunnen rangschikken (zonder herhaling van objecten) gelijk is aan

$$P(n, r) = \frac{n!}{(n-r)!} \tag{2.6}$$

waar we de notatie $P(n, r)$ introduceren.

We kunnen onze bevindingen ook veralgemenen naar een verzameling objecten waarin identieke objecten voorkomen.

Beschouw een verzameling van n objecten, waarvan n_1 van het eerste type, n_2 van het tweede type, ... en n_r van het r$^{\text{de}}$ type, met $\sum_{i=1}^{r} n_i = n$, dan bestaan er

$$\frac{n!}{n_1! n_2! n_3! ... n_r!} \tag{2.7}$$

mogelijke verschillende rangschikkingen van de n objecten.

[1] Het bewijs van de benaderingsformule van Stirling kan men vinden in andere cursussen van het eerste Bachelor jaar.

In een kaartspel [2] zijn er dus

$$\frac{52!}{13!\,13!\,13!\,13!} \tag{2.8}$$

mogelijke manieren om de kaarten te rangschikken volgens de 4 kleuren.

2.3 Combinaties

Na het begrip permutaties komen we tot de combinaties. Bij het oplossen van een typisch telprobleem moet men zich de vraag stellen of de *orde* van de rangschikking belangrijk is. De orde waarin de 6 balletjes van de nationale loterij getrokken worden, heeft geen enkel effect op je winstkansen, maar de orde van de gemakkelijke en moeilijke vragen op een examen zou wel eens belangrijk kunnen zijn en een invloed hebben op de slaagkansen (we beginnen dus meestal met de gemakkelijke !). Wanneer de orde inderdaad essentieel is, moeten we met permutaties en de productregel werken. Wanneer echter de orde van de rangschikking van de objecten geen effect heeft op het uiteindelijke resultaat, spreken we over combinaties.

Beschouw n objecten waaruit we een selectie of combinatie nemen van r objecten ($1 \leq r \leq n$) zonder naar de orde te kijken waarmee we de objecten nemen, dit is mogelijk op $C(n,r)$ manieren

$$C(n,r) = \left(\begin{array}{c} n \\ r \end{array} \right) = \frac{P(n,r)}{r!} = \frac{n!}{r! \cdot (n-r)!} \tag{2.9}$$

waar we de notatie $C(n,r)$ introduceren.

Op hoeveel manieren kan men bijvoorbeeld 6 balletjes trekken (zonder herhaling) uit een totaal van 42? Het antwoord is

$$C(42,6) = \frac{42!}{6!(42-6)!} = \frac{42 \cdot 41 \cdot 40 \cdot 39 \cdot 38 \cdot 37}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} \simeq 5.3 \cdot 10^6 \ . \tag{2.10}$$

Iemand die met de nationale lotto speelt, heeft dus weinig kans om te winnen. Let wel dat opeenvolgende trekkingen volledig onafhankelijk zijn en dat dezelfde sequentie van nummers terug dezelfde kans heeft. En ook dat elke sequentie een gelijke kans heeft.

Een labo-assistent moet zijn klas van 36 studenten opsplitsen in vier groepen van 9 studenten. Op hoeveel manieren kan hij die vier groepen maken? Voor de eerste groep kan de assistent 9 studenten kiezen uit 36, dus C(36,9). Voor de tweede groep blijven slechts 27 studenten over en hij moet er 9 uit kiezen, dus C(27,9). Dit gaat zo verder en we bekomen

$$C(36,9) \cdot C(27,9) \cdot C(18,9) \cdot C(9,9) = 2.145 \cdot 10^{19} \tag{2.11}$$

mogelijke manieren. Hopelijk komen de studenten dus overeen en organiseren ze zich zelf in vier groepen.

[2] Neem aan dat een kaartenspel 52 kaarten en 4 kleuren heeft: harten, ruiten, klaveren en schoppen (dit kan anders zijn voor een Chinees kaartspel).

Nu kunnen we enkele eenvoudige eigenschappen nagaan van de combinatie $C(n,r)$. Allereerst vinden we dat

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r! \cdot (n-r)!} = \frac{n!}{(n-r)! \cdot (n-(n-r))!} = \left(\begin{array}{c} n \\ n-r \end{array} \right) \tag{2.12}$$

en ook een twee eigenschap

$$\left(\begin{array}{c} n \\ r \end{array} \right) = \left(\begin{array}{c} n-1 \\ r \end{array} \right) + \left(\begin{array}{c} n-1 \\ r-1 \end{array} \right) \quad . \tag{2.13}$$

Bewijs van relatie 2.13: We willen een deelverzameling Y van r objecten kiezen uit een totale verzameling X van n objecten. Laat ons één willekeurig object x kiezen uit X (x∈X). Beschouw eerst dat object x geen deel uitmaakt van Y. Dan moeten we r objecten kiezen uit de overige n-1 objecten van de verzameling X, en dit kunnen we op C(n-1,r) mogelijke manieren. Beschouw anderszijds dat object x wel deel uitmaakt van Y. Dan moeten we nog r-1 elementen kiezen uit de overige n-1 objecten van de verzameling X, en dit kunnen we op C(n-1,r-1) mogelijke manieren. Omdat x wel of niet deel uitmaakt van de verzameling Y, moeten we beide mogelijkheden optellen. Dit is wat de vergelijking 2.13 weergeeft. QED.

De relatie 2.13 samen met de voorwaarden dat

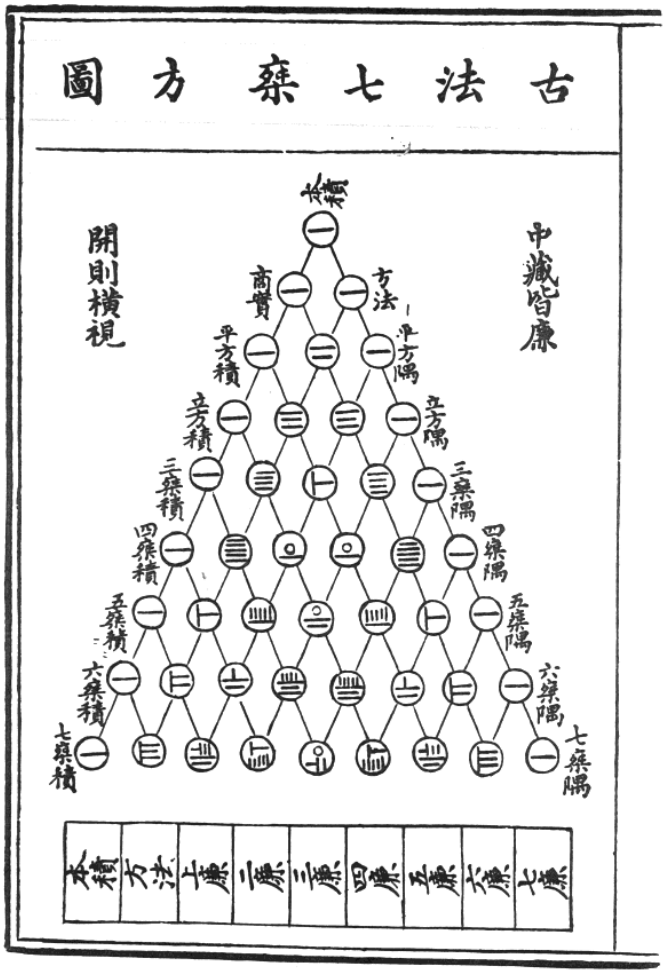
$$\left(\begin{array}{c} n \\ 0 \end{array} \right) = \left(\begin{array}{c} n \\ n \end{array} \right) = 1 \tag{2.14}$$

laat ons toe om de bekende driehoek van Pascal op te stellen. Deze driehoek geeft alle mogelijke waarden van $C(n,r)$ weer en berekent die via waarden die hogerop in de driehoek weergegeven zijn, zie Tabel 2.1.

C(n,r)	r=0	1	2	3	4	5	6	7	8	9	10
n=0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1

Tabel 2.1: *De driehoek van Blaise Pascal.*

Deze driehoek van Pascal vinden we reeds terug in de 14de eeuw in het Verre Oosten, zie Figuur 2.1.



Figuur 2.1: *De driehoek van Blaise Pascal zoals afgebeeld door Chu Shih-chieh rond het jaar 1303 verschillende eeuwen voor deze in de Westerse wereld opgesteld werd.*

We hebben gezien dat wanneer we uit een verzameling van n verschillende objecten r objecten willen rangschikken met mogelijke herhaling van eenzelfde object, we dit kunnen doen op n^r verschillende manieren. Nu willen we een gelijkaardig probleem beschouwen voor combinaties waar herhaling mogelijk is.

Beschouw het voorbeeld van 4 studenten die laat op de avond honger hebben en besluiten een frituur binnen te stappen. Elk van de studenten bestelt geen (en kiest bijvoorbeeld voor een drankje), één of meerdere van de volgende 'maaltijden': een mexicano, een bicky-burger, een pak friet met mayonaise of een pittabroodje. In totaal willen ze samen 7 items bestellen en ze laten de uitbater kiezen. De uitbater wil nagaan hoeveel mogelijkheden hij heeft om de studenten te bedienen. Hiervoor kan hij het probleem als volgt vereenvoudigen: hoe stop

je 7 balletjes in 4 verschillende dozen. Dit doet hij door de 7 balletjes op een rij te leggen en met 3 stokjes in totaal 4 categorieën te maken, bijvoorbeeld

$$xx|xx|xx|x \quad . \tag{2.15}$$

We hebben het probleem dus vereenvoudigd naar het rangschikken van 7+3=10 dingen waarvan 7 van de eerste soort en 3 van de tweede soort. Dit geeft in totaal

$$\frac{(7+3)!}{7! \cdot 3!} = \frac{(7+(4-1))!}{7! \cdot (4-1)!} \tag{2.16}$$

mogelijkheden aan de uitbater om de studenten hun honger te stillen.

Algemeen kunnen we stellen dat indien we uit een verzameling van n verschillende objecten er r willen selecteren met mogelijke herhaling van hetzelfde object, we in totaal

$$C(n+r-1,r) = \left(\begin{array}{c} n+(r-1) \\ r \end{array} \right) = \frac{(n+r-1)!}{r! \cdot (n-1)!} \tag{2.17}$$

mogelijkheden hebben. Dit komt neer op r objecten (x) met $(n-1)$ onderscheidingen ($|$).

2.4 Binomiaalwet

Stel dat we vier muntstukken opgooien. Voor elk muntstuk is de waarschijnlijkheid om 'kop' te bekomen gelijk aan 1/2 en om 'munt' te bekomen gelijk aan 1/2. Dit is een typisch experiment met slechts twee mogelijke uitkomsten, ofwel 'succes' of 'geen succes', ofwel 'raak' of 'mis', ofwel 'observatie' of 'geen observatie' van een gebeurtenis, enzovoort. Dergelijke experimenten noemen we Bernoulli *trial* processen: het achtereenvolgens uitvoeren van n kansexperimenten zodat elk experiment twee mogelijke uitkomsten heeft, de ene met een waarschijnlijkheid p, de andere met een waarschijnlijkheid $q = 1 - p$. Ook moet gelden dat de uitkomst van experiment n niet beïnvloed wordt door de $n-1$ voorgaande.

Het experiment met de vier muntstukken heeft een aantal mogelijke uitkomsten die we even in detail opsommen:

- Alle muntstukken komen neer als 'kop'. Vermits we aannemen dat het effect van het opgooien van een muntstuk onafhankelijk is van elk ander muntstuk, wordt de totale waarschijnlijkheid van deze uitkomst bepaald door de productregel en is dus gelijk aan $(1/2)^4$, we noteren dit met P(4);

- Drie muntstukken komen neer als 'kop' en bijgevolg één als 'munt'. De waarschijnlijkheid om dit te bekomen is terug $(1/2)^4$ als we willen dat het laatste muntstuk 'munt' is. Als we niet geïnteresseerd zijn in de volgorde van de uitkomsten, kunnen we 4 mogelijke sequenties bekomen: KKKM, KKMK, KMKK en MKKK. Elk van deze heeft een waarschijnlijkheid van $(1/2)^4$ van voorkomen en samen heeft deze uitkomst bijgevolg een waarschijnlijkheid van $P(3) = 4 \cdot (1/2)^4$;

- Twee muntstukken komen neer als 'kop' en bijgevolg twee als 'munt'. Hiervoor zijn er zes verschillende sequenties mogelijk: MMKK, MKMK, MKKM, KKMM, KMKM en KMMK. De totale waarschijnlijkheid voor deze uitkomst is bijgevolg gelijk aan $P(2) = 6 \cdot (1/2)^4$;

- Eén muntstuk komt neer als 'kop' en de drie andere als 'munt'. De waarschijnlijkheid voor deze uitkomst is analoog aan de waarschijnlijkheid *drie muntstukken komen neer als 'kop' en één als 'munt'*. Bijgevolg is de waarschijnlijkheid $P(1) = 4 \cdot (1/2)^4$;

- Alle muntstukken komen neer als 'munt'. Dit is terug analoog met de uitkomst *alle muntstukken komen neer als 'kop'* en de waarschijnlijkheid is bijgevolg gelijk aan $P(0) = (1/2)^4$.

Door rekening te houden met het axioma dat de totale waarschijnlijkheid voor alle uitkomsten samen gelijk moet zijn aan 1, kunnen we nagaan of we alle mogelijke uitkomsten hebben

$$\sum_{i=0}^{4} P(i) = P(0) + P(1) + P(2) + P(3) + P(4) = \frac{16}{16} = 1 \quad . \tag{2.18}$$

Dit voorbeeld kunnen we veralgemenen naar een Bernoulli experiment met twee mogelijke uitkomsten A en $\overline{A}$, de ene met een waarschijnlijkheid p en de andere bijgevolg met een waarschijnlijkheid $q = 1 - p$. We stellen $P(i)$ gelijk aan de waarschijnlijkheid dat bij een totaal van n waarnemingen juist i maal uitkomst A voorkomt. In dit geval bekomen we i maal uitkomst A en $(n - i)$ maal uitkomst $\overline{A}$, bijgevolg is de waarschijnlijkheid

$$p^i \cdot q^{n-i} = p^i \cdot (1 - p)^{n-i} \tag{2.19}$$

indien de volgorde waarin A en $\overline{A}$ optreden gegeven is. Is deze volgorde willekeurig, dan kan men i herhalingen van A op $C(n, i)$ mogelijke manieren verkrijgen, zodat we voor de waarschijnlijkheid de volgende uitdrukking bekomen

$$P(i) = C(n, i) \cdot p^i \cdot (1 - p)^{n-i} = \left(\begin{array}{c} n \\ i \end{array} \right) \cdot p^i \cdot (1 - p)^{n-i} \quad . \tag{2.20}$$

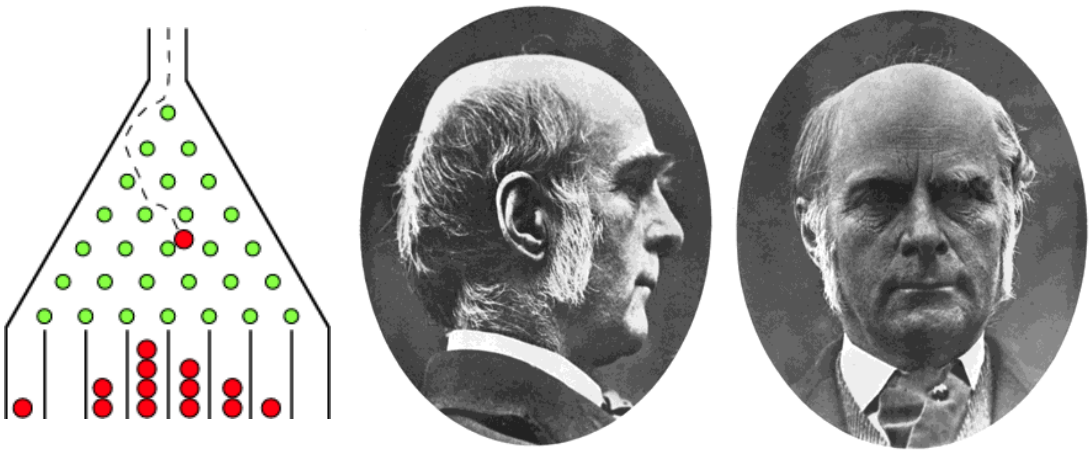
De uitdrukking 2.20 noemen we de binomiaalwet die soms ook onder de volgende vorm geschreven wordt

$$(x + y)^n = \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) \cdot x^i \cdot y^{n-i} = \sum_{i=0}^{n} \left(\begin{array}{c} n \\ n - i \end{array} \right) \cdot x^i \cdot y^{n-i} \quad . \tag{2.21}$$

waar de laatste gelijkheid geldt door de symmetrie van de uitdrukking. De coëfficiënten $C(n, i)$ noemen we gewoonlijk de binomiaalcoëfficiënten die we via de driehoek van Pascal eenvoudig kunnen bepalen.

Een educatieve toepassing van de binomiaalwet vinden we terug in het bord van Galton, zie Figuur 2.2. Sir Francis Galton (1822-1911) was een Engelse wetenschapper en was een neef van de welgekende Charles Darwin. Een balletje begint bovenaan het bord en heeft aan elk stokje twee mogelijkheden: 'links' of 'rechts'. Indien het bord goed gemaakt is, dan heeft het balletje bij ieder stokje een gelijke kans om 'links' of 'rechts' te vallen. Ook is

iedere keuze bij een stokje op laag i onafhankelijk van de $i-1$ vorige keuzes. Elk balletje dat onderaan komt, heeft dus in totaal n Bernoulli experimenten gedaan (n = aantal lagen) en bijgevolg zal elke aankomstplaats een waarschijnlijkheid hebben gedefinieerd volgens de binomiaalcoëfficiënten $C(n,i)$. Let op de sterke gelijkenis met de 'driehoek' van Pascal !!



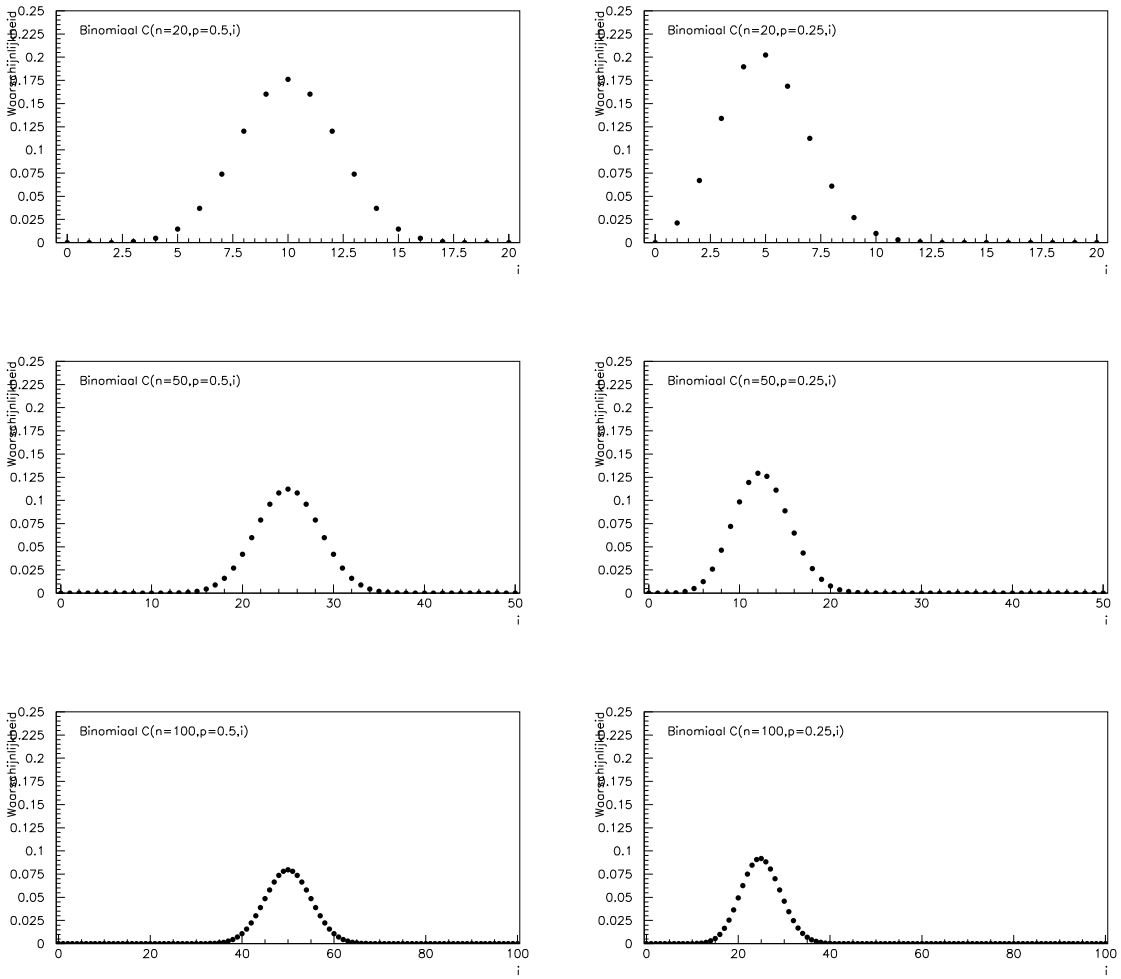
Figuur 2.2: *Het bord van Galton als illustratie van de binomiaalwet en enkele foto's van Sir Francis Galton.*

Hoeveel balletjes er zich in de individuele hokjes onderaan het bord van Galton bevinden, is afhankelijk van het aantal lagen n (= het aantal opeenvolgende identieke Bernoulli experimenten) en de kans p om 'rechts' te bekomen als uitkomst (= wat eigenlijk de plaats van het hokje aangeeft). Om een idee te hebben hoe de balletjes verdeelt zijn over de hokjes voor verschillende waarden van de parameters n en p, kunnen we dit experiment met de computer simuleren [3].

In Figuur 2.3 worden de waarschijnlijkheden $P(i)$ weergegeven voor enkele parameters. De linkse grafieken tonen de waarschijnlijkheden om een balletje terug te vinden in elk hokje onderaan het bord van Galton en dit voor een verschillend aantal lagen n, namelijk 20, 50 en 100 (bijgevolg ook een verschillende aantal hokjes, $n+1$). De grafieken die rechts staan tonen dezelfde waarschijnlijkheden, maar dit maal voor een bord van Galton waar de kans om 'links' te bekomen bij elk Bernoulli experiment gelijk is aan 0.7 ('rechts' heeft bijgevolg een kans van 0.3).

Dit voorbeeld motiveert het gebruik van een variabele X die een uitkomst van een experiment is en die waarden kan aannemen in de steekproefruimte Ω. We gaan hiermee verder in het volgend hoofdstuk.

[3] Op het internet kan je dergelijke simulatie vinden op bijvoorbeeld volgende website: http://stat-www.berkeley.edu/~stark/Java/BinHist.htm



Figuur 2.3: *Grafi sche voorstellingen van de waarschijnlijkheid dat een balletje dat bovenaan het bord van Galton begint, onderaan in een zeker hokje i terecht komt.*

Hoofdstuk 3

Stochastische variabelen en verdelingen

"La théorie des probabilités n'est que
le bon sens reduit au calcul"

P.-S. le Marquis de Laplace,
Mécanique Céleste, 1799

In vorige hoofdstukken hebben we gezien hoe we een waarschijnlijkheid definiëren en hoe we ermee moeten rekenen. Om de empirische uitkomsten van een experiment eenvoudig te kwantificeren gaan we in dit hoofdstuk enkele begrippen of grootheden invoeren. Deze begrippen kunnen we ook toepassen op experimenten die we hypothetisch oneindig keer herhalen, dan spreken we over theoretische grootheden in plaats van empirische grootheden.

3.1 Definitie

Het uitvoeren van een wetenschappelijk experiment wordt beïnvloed door verschillende factoren. Als we willen weten hoeveel liter brandstof onze auto verbruikt per kilometer, kunnen we dat door eenvoudig het aantal gereden kilometers te delen door het aantal verbruikte liters. Als we dit experiment verschillende keren herhalen, zullen we echter nooit hetzelfde reële getal uitkomen. Er zijn tal van factoren die onze uitkomst beïnvloeden: hebben we het aantal kilometers en de hoeveelheid brandstof juist gemeten, was er veel verkeer op de baan of weinig, hadden we tegenwind, reden we in Nederland op de vlakke weg of bergop in Oostenrijk, hebben we veel moeten remmen, enzovoort. Het is onmogelijk om al deze factoren in rekening te brengen tijdens het uitvoeren van ons experiment. We kunnen bijgevolg nooit exact de uitkomst voorspellen, of anders gezegd een welbepaalde waarneming is nooit exact reproduceerbaar. De verschillende waarnemingen of uitkomsten van een experiment vertonen een spreiding of variatie.

We definiëren iedere waarnemingsgrootheid waarvan de waarde op een dergelijke wijze varieert als een toevallige of stochastische veranderlijke. Dit impliceert dat voor elk herhaalbaar experiment met de stochastische variabele X als uitkomst, die de betrachting heeft een

fysische grootheid μ met de juiste en constante waarde μ_0 te meten, men tracht uit de verzameling van bekomen resultaten $\{x_1, x_2, ..., x_n\}$ een conclusie te formuleren in verband met de waarde van de fysische grootheid. Omdat het experiment beïnvloed wordt door verschillende willekeurige of random factoren is het verband tussen de uitkomsten $\{x_1, x_2, ..., x_n\}$ en de echte constante waarde van μ, namelijk μ, niet triviaal. Men tracht met verschillende conclusies over de waarde van μ verschillende waarschijnlijkheden te associëren.

Om het mogelijk te maken om via de uitkomsten $\{x_1, x_2, ..., x_n\}$ iets te concluderen over de echte waarde van μ, moeten de experimenten *at random* genomen zijn. Daar het onmogelijk is om oneindig veel experimenten uit te voeren, moet men een steekproef nemen. Deze steekproef van experimenten moet een zo goed mogelijke weerspiegeling zijn van de oneindig grote steekproefruimte die we niet kunnen opmeten. Bij het bepalen van het brandstofverbruik per kilometer van onze auto moeten we ervoor zorgen dat we de verbruikte brandstof niet steeds opmeten als we bergop rijden of in de file staan. Dit zou namelijk een verkeerd beeld geven van het totale verbruik van onze auto.

- **Discrete variabelen**

  Indien we een experiment uitvoeren waarvan we de uitkomst moeten zoeken in een steekproefruimte waarvan het aantal elementen telbaar is (= eindig aantal), dan associëren we met de uitkomst van het experiment een discrete stochastische variabele. Het gooien van een dobbelsteen bijvoorbeeld heeft 6 discrete uitkomsten.

- **Continue variabelen**

  Een experiment waarvan de mogelijke uitkomsten een oneindige verzameling van elementen omvat, wordt gekarakteriseerd met een continue stochastische variabele. Het resultaat van een meting naar het brandstofverbruik per kilometer van een auto heeft een continue uitkomst, elk reëel getal is mogelijk.

Door het uitvoeren van experimenten bekomen we een verzameling gegevens of uitkomsten $\{x_1, x_2, ..., x_n\}$. Deze empirische gegevens trachten we te beschrijven via grootheden, bekomen uit deze verzameling gegevens.

- **Gemiddelde**: De meest voorkomende grootheid waarmee de ligging van de data gekarakteriseerd wordt, is het rekenkundig gemiddelde $\overline{x}$. Het gemiddelde van een steekproef van n metingen is:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \quad . \tag{3.1}$$

- **Aantal gegevens**: Het aantal uitgevoerde experimenten n is een belangrijke grootheid die (zoals we later zullen zien) informatie geeft over hoe nauwkeurig we bijvoorbeeld het gemiddelde van een verzameling gegevens kunnen bepalen. Indien we het gemiddelde berekend hebben wanneer n groot is, zal een bijkomend of $n+1^{de}$ experiment de waarde van het gemiddelde niet veel meer veranderen. We kunnen bijgevolg concluderen dat het gemiddelde goed gekend is (met een kleine 'onzekerheid'). Dit zou niet het geval zijn indien we een gelijkaardig experiment slechts $n/10$ keer uitvoeren. Het

gemiddelde zou dan relatief minder goed gekend zijn en dus meer beïnvloed worden door een bijkomende waarneming.

- **Mediaan**: Het rekenkundig gemiddelde is een wiskundig belangrijke grootheid om theoretische stochastische verdelingen te beschrijven, maar is bij een eindig aantal metingen heel gevoelig aan mogelijke uitschieters (*outliers* in het Engels). Het gemiddelde van een verzameling van 10 metingen met waarden $\{1, 1, 1, 1, 1, 1, 1, 1, 1, 10\}$ geeft 1.9 als resultaat. De laatste meting met uitkomst 10 heeft bijgevolg een enorme invloed op het gemiddelde en kan men beschouwen als een uitschieter (het gemiddelde zou een totaal andere waarde hebben indien deze meting niet in rekening genomen wordt). Een meer robuuste grootheid is de mediaan die gedefinieerd wordt als de 'middelste' waarneming, waarbij we het concept 'middelste' interpreteren als het gemiddelde van de twee middelsten als het aantal waarnemingen even is. Om de mediaan te bepalen moeten we bijgevolg de verschillende metingen rangschikken of ordenen. De mediaan is de waarde voor dewelke 50% van de metingen een kleinere uitkomst heeft en 50% een grotere uitkomst. Voor ons voorbeeld van 10 metingen zal de mediaan die de waarde 1 heeft, niet veranderen indien we de laatste meting al dan niet in rekening brengen.

- **Standaardafwijking**: De meest voorkomende grootheid waarmee de schaal van de gegevens gekarakteriseerd wordt, is de standaardafwijking of spreiding s. Een alternatief voor de standaardafwijking is de steekproef variantie Var die gedefinieerd wordt als het kwadraat van de standaardafwijking

$$\mathrm{Var} = s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{3.2}$$

  waar $\overline{x}$ het rekenkundig gemiddelde is. Het voordeel van de grootheid s is dat die dezelfde dimensies heeft als het gemiddelde en dus ook grafisch op dezelfde as voorgesteld kan worden. De reden waarom we door $(n-1)$ delen in plaats van n zal later duidelijk worden (zie Hoofdstuk 7).

- **Percentielen**: In de praktijk is het soms nuttig om te weten voor welke waarde van x in totaal α% van de metingen kleiner dan of gelijk zijn aan x. Dit geven we weer aan de hand van α-percentielen. Met α tussen 0% en 100% definiëert het α-percentiel het punt x waaronder in totaal α% van de empirische metingen liggen. De mediaan is bijgevolg het 50%-percentiel van de metingen.

Dit zijn empirische grootheden bekomen uit de gegevens van een eindig aantal experimenten. Indien men verschillende verzamelingen van gegevens wil vergelijken, is het soms nuttig om de verzameling gegevens samen te vatten met een aantal grootheden. We kunnen bijvoorbeeld zien of de twee verzamelingen hetzelfde gemiddelde en bijgevolg dezelfde ligging hebben. Alsook kunnen we de dispersie of spreiding van de gegevens vergelijken met behulp van de standaardafwijking. In de praktijk echter gaat men de empirische gegevens veelal benaderen door een theoretische verdeling van de waarschijnlijkheid, en de parameters van deze curve publiceren alszijnde de resultaten. We komen hierop terug in hoofdstuk 7.

3.2 Waarschijnlijkheidsverdelingen

In de hypothese dat we oneindig veel experimenten kunnen uitvoeren, beschikken we bijgevolg over een oneindige verzameling uitkomsten. Deze uitkomsten kunnen we enkel analytisch beschrijven via een functie f_X die de waarschijnlijkheid voor elke uitkomst weergeeft. Dergelijke functies noemen we waarschijnlijkheidsdichtheidsverdelingen. Let erop dat dit theoretische concepten zijn, daar we nooit oneindig veel experimenten kunnen uitvoeren. De experimentele gegevens die we verzamelen, kunnen hoogstens het patroon vertonen van dergelijke theoretische waarschijnlijkheidsdichtheidsverdeling. In volgende hoofdstukken zullen we zien dat, indien we veel experimenten uitvoeren en bijgevolg een grote verzameling uitkomsten voor een stochastische variabele bekomen, we de eigenschappen van de theoretische verdelingen kunnen toekennen aan de empirische gegevens.

De waarschijnlijkheidsdichtheidsverdeling f_X wordt gedefinieerd als de afgeleide van de cumulatieve verdelingsfunctie F_X. Als X een stochastische variabele is, dan is de functie F_X, die voldoet aan $F_X(a) = P(X \leq a)$, de cumulatieve verdelingsfunctie van X. Deze functie $F_X(a)$ neemt bijgevolg waarden aan tussen 0 en 1 ($0 \leq F_X(a) \leq 1$), en geeft de waarschijnlijkheid weer dat de uitkomst van een experiment met stochastische variabele X kleiner is dan a.

We noemen een stochastiek X discreet als X slechts een eindig (of aftelbaar oneindig) aantal verschillende waarden kan aannemen. Dat wil zeggen dat er een verzameling reële getallen $\{x_i \mid i = 1, 2, ..., n\}$ is, zodat $P(X = x_i) = p_i$ en dat de totale kans 1 blijft, namelijk $\sum_{i=1}^{n} p_i = 1$.

We noemen een stochastische variabele X continu als de cumulatieve verdelingsfunctie F_X een continue en overal differentieerbare functie is (behalve in een eindig aantal punten). Dit laat toe om de waarschijnlijkheidsdichtheidsverdeling f_X te definiëren als de afgeleide van F_X, of

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{3.3}$$

en bijgevolg ook omgekeerd, de cumulatieve verdelingsfunctie te definiëren als de integraal van de waarschijnlijkheidsdichtheidsverdeling, of

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt \ . \tag{3.4}$$

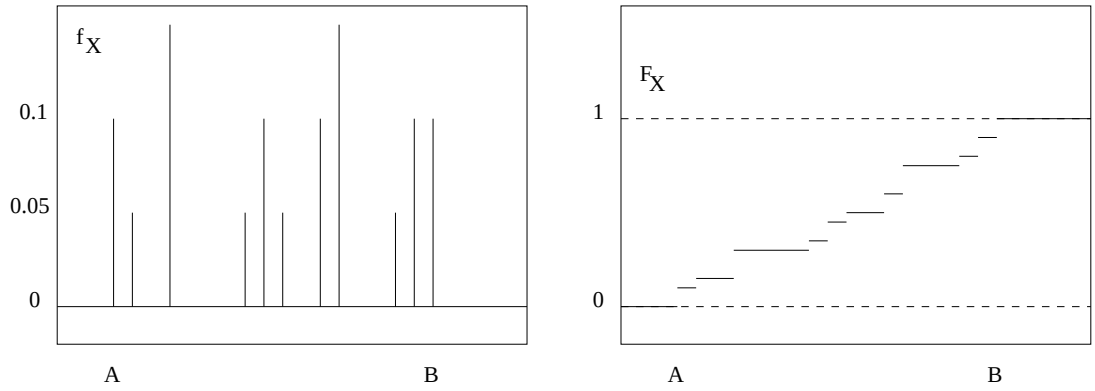
In Figuur 3.1 vinden we een voorbeeld van een discrete kansverdeling f_X en bijhorende cumulatieve verdeling F_X, in Figuur 3.2 idem voor een continue kansverdeling.

Een vereiste voor zowel de discrete als de continue verdelingen is dat die genormaliseerd moeten zijn, de totale waarschijnlijkheid moet namelijk 1 zijn. Voor de discrete verdelingen komt dit neer op het volgende

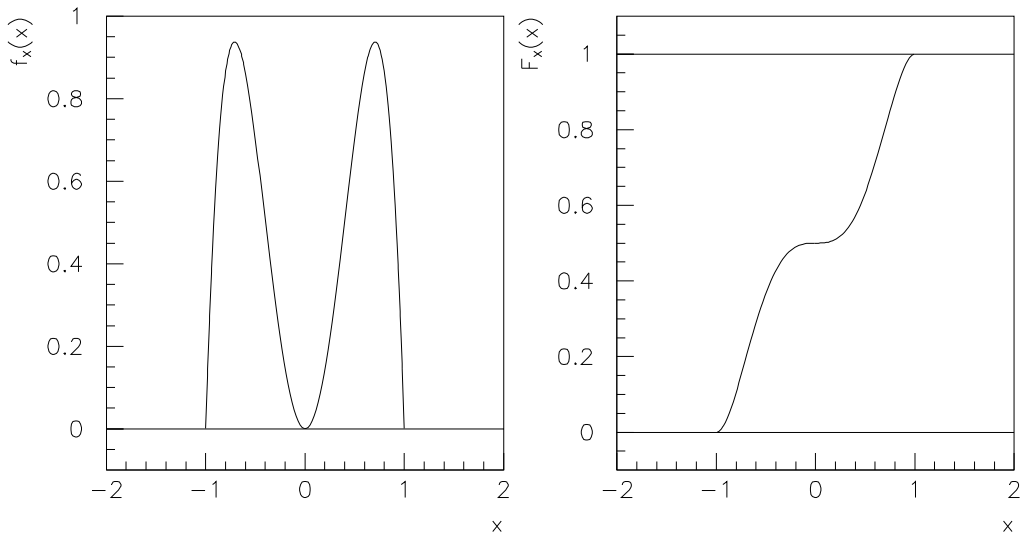
$$\sum_{i=1}^{n} f_X(x_i) = 1 \tag{3.5}$$

en voor de continue verdelingen moet het volgende gelden

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1 \ . \tag{3.6}$$



Figuur 3.1: *Illustratie van de verdelingen f_X en F_X van een discrete stochastische variabele.*



Figuur 3.2: *Illustratie van de verdelingen f_X en F_X van een continu stochastische variabele.*

Indien deze voorwaarde niet geldt wanneer een theoretische verdeling $f_X(x)$ gegeven wordt, dan moet een normalisatiefactor C_{norm} ingevoerd worden. De verdelingen moeten dan herschaald worden met een factor C_{norm} die onafhankelijk is van de stochastiek X en die de verhoudingen van de waarschijnlijkheden behoudt. Dit gebeurt op volgende manier voor een discrete verdeling

$$f_X^{norm}(x_i) = C_{norm} \cdot f_X(x_i) \quad \forall i \in \{1, 2, ..., n\} \tag{3.7}$$

en voor een continue verdeling

$$f_X^{norm}(x) = C_{norm} \cdot f_X(x) \tag{3.8}$$

waar de waarde voor C_{norm} bepaald wordt als volgt

$$C_{norm} = \frac{1}{\sum_{i=1}^{n} f_X(x_i)} \tag{3.9}$$

voor een discrete verdeling, en

$$C_{norm} = \frac{1}{\int_{-\infty}^{+\infty} f_X(x)dx} \tag{3.10}$$

voor een continue verdeling. Voor de genormaliseerde verdelingen $f_X^{norm}(x)$ geldt dan wel de relatie 3.5 of 3.6.

Ook voor de theoretische verdelingen kunnen we kentallen definiëren. Dit zijn dan theoretische grootheden, bekomen uit de analytische voorstelling van een statistische variabele, namelijk het functievoorschrift. Deze wiskundige functies worden gekarakteriseerd door enkele constanten die we parameters noemen. Neem bijvoorbeeld de klassieke Maxwell-Boltzmann [1] verdeling:

$$f(E) = \frac{1}{Ae^{E/kT}} \tag{3.11}$$

welke de waarschijnlijkheidsdichtheid weergeeft dat een deeltje een zekere energie E bezit. In deze verdeling vinden we twee constante parameters, namelijk de constante van Boltzmann k (gelijk aan 1.3807 10^{-23} J/K) en de temperatuur T (die ook voorkomt in de normalisatieconstante A). Indien één van beide parameters verandert, hebben we een nieuwe waarschijnlijkheidsdichtheidsverdeling gedefinieerd.

Als we verschillende verdelingen willen vergelijken, moeten we de analytische functie grafisch voorstellen en met onze eigen ogen de bekomen grafieken vergelijken. Dit is duidelijk een process dat weinig kwantitatief is. Daar we hier geen experimenten uitgevoerd hebben, maar uitgaan van een hypothetische verzameling van oneindig veel experimenten, moeten we spreken over theoretisch verwachtte grootheden in plaats van empirisch bepaalde grootheden. Bijgevolg moeten we een functie $E[X]$ definiëren [2] die gebruik maakt van de stochastiek X en een verwachtingswaarde als functiewaarde geeft. Als $g(X)$ een functie is van stochastiek X met waarschijnlijkheidsdichtheidsverdeling f_X, dan is de verwachtingswaarde van $g(X)$ voor een discrete verdeling gelijk aan

$$E[g(X)] = \sum_{i=1}^{n} g(x_i)\ f_X(x_i) \tag{3.12}$$

en voor een continue verdeling gelijk aan

$$E[g(X)] = \int_{\Omega} g(x)\ f_X(x)\ dx \tag{3.13}$$

[1] We gaan ervan uit dat een oneindig aantal deeltjes zich in eenzelfde energietoestand kunnen bevinden.

[2] De notatie $E[X]$ komt uit het Engels waar 'E' staat voor *Expectation*.

waar Ω de steekproefruimte van X is. Merk op dat de verwachtingswaarde niet steeds bestaat, daar de reeks of de oneindige integraal kan divergeren. Er bestaan eenvoudige rekenregels voor de functie $E[X]$

- $E[aX] = a\ E[X]$, voor elke $a \in \Re$ met $a \neq 0$;
- $E[X + b] = E[X] + b$, voor elk $b \in \Re$;
- $E[b] = b$, voor elke $b \in \Re$;
- $|E[X]| \leq E[|X|]$.

Dit brengt ons tot het definiëren van de kentallen of de momenten van een theoretische verdeling. Het k^{rmde} orde moment van de verdeling van stochastiek X definiëren [3] we als ($k \in \{0, 1, 2, ...\}$)

$$\mu_k(X) = E[X^k] = \int_{-\infty}^{+\infty} x^k\ f_X(x)\ dx \tag{3.14}$$

of de verwachtingswaarde van de verdeling X^k. Deze is bijgevolg zelf geen functie meer van x, enkel een theoretische grootheid die de verdeling van de stochastiek X karakteriseert. De gemiddelde waarde van een verdeling is het moment van eerste orde en noteren we als $\mu = \mu_1$. Het centrale moment van de verdeling van stochastiek X van orde k definiëren we als

$$\mu'_k(X) = E[(X - E[X])^k] = E[(X - \mu)^k] = \int_{-\infty}^{+\infty} (x - \mu)^k\ f_X(x)\ dx \tag{3.15}$$

of de verwachtingswaarde van de verdeling $(X - \mu)^k$. Daar de verwachtingswaarde van sommige verdelingen niet bestaat, zullen ook sommige verdelingen geen momenten van een bepaalde orde hebben. Wel geldt dat indien het r^{de} moment van een verdeling bestaat, ook alle lagere orden bestaan.

Via de binomiaalwet kunnen we een verband leggen tussen de momenten en de centrale momenten van een verdeling, namelijk

$$(x - \mu)^k = x^k - \begin{pmatrix} k \\ 1 \end{pmatrix}\ x^{k-1}\mu + \begin{pmatrix} k \\ 2 \end{pmatrix}\ x^{k-2}\mu^2 - ... + (-1)^k\ \mu^k\ . \tag{3.16}$$

Nemen we van beide leden de verwachtingswaarde, dan bekomen we

$$\mu'_k = \mu_k - \begin{pmatrix} k \\ 1 \end{pmatrix}\ \mu_{k-1}\ \mu + \begin{pmatrix} k \\ 2 \end{pmatrix}\ \mu_{k-2}\ \mu^2 - ... + (-1)^k\ \mu^k \tag{3.17}$$

en daar $\mu_0 = 1$ en $\mu_1 = \mu$ kunnen we een verband tussen beide types momenten opstellen.

Met behulp van deze momenten kunnen we het gemiddelde en de standaardafwijking definiëren van een theoretische verdeling. Zoals reeds vermeld, kunnen we het gemiddelde

[3] Analoge definitie voor discrete verdelingen.

van de verdeling bepalen als het eerste orde moment μ_1, terwijl de theoretisch variantie $\mathrm{Var}[X]$ van de verdeling van de stochastische variabele X gegeven wordt door

$$\mathrm{Var}[X] = E[(X - E[X])^2] = \mu_2'(X) \tag{3.18}$$

het tweede centrale moment, of

$$\mathrm{Var}[X] = E[X^2] - (E[X])^2 = \mu_2 - \mu_1^2 \tag{3.19}$$

een combinatie van gewone momenten. In sommige toepassingen is het eenvoudiger om de eerste uitdrukking te gebruiken, terwijl voor andere toepassingen de tweede eenvoudiger is. De theoretische standaardafwijking σ_x bekomt men via [4]

$$\sigma_x = \sqrt{\mathrm{Var}[X]} \ . \tag{3.20}$$

Hiermee kunnen we ook enkele nuttige eigenschappen aantonen van de variantie, namelijk

- $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$
- $\mathrm{Var}[X + b] = \mathrm{Var}[X]$
- $\mathrm{Var}[b] = 0$

Het concept van momenten van verdelingen kunnen we uiteraard ook toepassen op empirische gegevens in plaats van op theoretische verdelingen. We moeten dan de integralen voor de continue verdelingen vervangen door sommaties over de $n < \infty$ verschillende metingen.

Met behulp van de momenten willen we ook de vorm van de verdeling, namelijk de symmetrie, karakteriseren. Onder de kentallen die de vorm definiëren van een waarschijnlijkheidsdichtheidsverdeling vinden we de zogenaamde scheefheid (in het Engels de *skewness*) en kurtosis.

Voor een symmetrische verdeling vallen het gemiddelde en de mediaan samen. Een grootheid die de symmetrie van een verdeling karakteriseert zou bijgevolg het verschil kunnen zijn tussen het rekenkundig gemiddelde en de mediaan. Ook kunnen we aantonen dat voor een symmetrische verdeling elk centraal moment van oneven orde gelijk is aan nul. Bijgevolg kunnen we bijvoorbeeld het derde centrale moment μ_3 als een maat van de scheefheid van een verdeling beschouwen. Om uiteindelijk een dimensieloze grootheid te bekomen, delen we μ_3 door de derde macht van de standaardeviatie en definiëren we de scheefheidscoëfficiënt γ_1 van stochastiek X als

$$\gamma_1(X) = \frac{\mu_3}{\sigma_x^3} = \frac{\mu_3}{\mu_2^{3/2}} \ . \tag{3.21}$$

Indien $\gamma_1(X)$ verschillend is van nul, kunnen we concluderen dat de verdeling niet symmetrisch is.

[4] De waarde van de grootheden aangeduid met Griekse symbolen μ_x en σ_x halen we uit theoretische verdelingen, terwijl we de waarde van grootheden aangeduid met Latijnse symbolen $\overline{x}$ en s_x uit empirische gegevens halen.

Een ander kenmerk van de vorm van een verdeling is de dikte van de 'staarten' van de verdeling. Als de uiteinden van de verdeling een relatief grote dichtheid hebben, vergeleken met het midden van de verdeling, spreken we van dikke of dichtbevolkte staarten. Dit kwantificeren we met behulp van een vierde centrale moment, die door de factor $(x - E[X])^4$ in zijn uitdrukking relatief meer bijdrage zal hebben van die staarten vergeleken met de lagere momenten zoals het tweede centrale moment. Om alweer een dimensieloze grootheid te bekomen, delen we door de vierde macht van de standaardafwijking en bekomen we

$$\gamma_2(X) = \frac{\mu_4}{\sigma_x^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3 \tag{3.22}$$

als uitdrukking voor de zogenaamde kurtosis (Grieks voor welving). Het getal 3 komt overeen met de waarde van $\frac{\mu_4}{\mu_2^2}$ voor de normale verdeling (zie Hoofdstuk 4). Men bepaalt bijgevolg de dichtheid van de staarten van de verdeling relatief ten opzichte van deze van de normale verdeling. Een verdeling met positieve $\gamma_2(X)$ heeft dichtbevolkte staarten (ook wel platycurtic genoemd omdat men initieel verkeerd dacht dat dit verdelingen zijn met een platte top), terwijl een negatieve waarde voor $\gamma_2(X)$ aanduidt dat de staarten dunbevolkt zijn (ook wel leptocurtic genoemd).

3.3 Meer-dimensionale verdelingen

In vele gevallen kunnen we aan de elementen in de steekproefruimte Ω meerdere stochastische variabelen toekennen. In het universum vinden we onder andere een verzameling sterren. We kunnen deze beschouwen als elementen $a \in \Omega$ van de steekproefruimte Ω en aan elk element verschillende stochatische variabelen $\{X_i \mid i \in \{1, 2, ..., n\}\}$ toekennen. Deze variabelen kunnen bijvoorbeeld zijn: coördinaten ten opzichte van de Aarde, temperatuur, straal, kleur, druk in het centrum, enzovoort. Dit kan bijvoorbeeld ook het gelijktijdig werpen van twee dobbelstenen zijn, waar X de stochastische variabele is die de uitkomsten van de ene dobbelsteen beschrijft en Y de stochastische variabele voor de andere dobbelsteen. Elk van deze variabelen X_i kan waarden x_i aannemen die zowel discreet als continu kunnen zijn. We spreken dan van een n-dimensionale stochastische variabele of kansvector.

We beschouwen hier enkel het speciale geval van twee-dimensionale veranderlijken $z = (x, y)$. Hier moeten we de veranderlijke Z interpreteren als coördinaten in het vlak, waarbij x de waarde van de coördinaat geeft langs één as en y de waarde van de coördinaat langs een tweede as. De cumulatieve verdelingsfunctie F_Z van een twee-dimensionale kansvector $Z = (X, Y)$ wordt gedefinieerd als volgt

$$F_Z(a, b) = P(X \leq a \quad \text{en} \quad Y \leq b) \ . \tag{3.23}$$

Zoals in het vorige kunnen we alweer een onderscheid maken tussen discrete en continue kansvectoren. Om de definitie van de waarschijnlijkheidsdichtheidsverdeling mogelijk te maken voor continue kansvectoren, leggen we de extra eis op dat de gemengde tweede afgeleide van de cumulatieve verdeling F_Z moet bestaan, zodat

$$f_Z(x, y) = \frac{\partial^2 F_Z(x, y)}{\partial x \partial y} \tag{3.24}$$

en omgekeerd

$$F_Z(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_Z(u,v)\,dudv \tag{3.25}$$

waar $f_Z(u,v)dudv$ de waarschijnlijkheid weergeeft dat x een waarde zou aannemen in het oneindig klein interval $(u, u+du)$ en dat y een waarde zou aannemen in het oneindig klein interval $(v, v+dv)$. Bijgevolg kunnen we voor iedere deelverzameling A van de steekproefruimte Ω de waarschijnlijkheid bepalen dat een experiment resulteert in de waarden (x,y) met $z=(x,y)\in A$, namelijk

$$P(Z\in A) = \int\int_A f_Z(u,v)\,dudv \ . \tag{3.26}$$

of ook

$$P(Z\in A) = F_Z(u_2,v_2) - F_Z(u_1,v_2) - F_Z(u_2,v_1) + F_Z(u_1,v_1) \tag{3.27}$$

indien A afgebakend wordt door de coördinaten $u_1 < x \leq u_2$ op de eerste as en door de coördinaten $v_1 < y \leq v_2$ op de tweede as. Als we de integralen vervangen door sommaties, kunnen we op een gelijkaardige manier de waarschijnlijkheid bepalen in een discrete steekproefruimte. Alweer moet gelden dat de totale waarschijnlijkheid 1 is of

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_Z(u,v)\,dudv = 1 \ . \tag{3.28}$$

Indien deze vergelijking niet geldt, moet men alweer constante normalisatiefactoren C_{norm} invoeren die onafhankelijk zijn van de waarden x en y. Men vermenigvuldigt de waarschijnlijkheidsdichtheid $f_Z(x,y)$ met eenzelfde factor C_{norm} die constant is voor elke waarde x en y.

Uitgaande van de twee-dimensionale verdeling $f_Z(x,y)$ kunnen we de één-dimensionale verdelingen $f_X(x)$ en $f_Y(y)$ bekomen. Indien we de verdeling $f_X(x)$ willen bekomen, moeten we veronderstellen dat y elke mogelijke waarde in de steekproefruimte van Y kan aannemen. We moeten dus integreren over alle mogelijke waarden van y of

$$f_X(x) = \int_{-\infty}^{+\infty} f_Z(x,y)\,dy \tag{3.29}$$

en analoog voor de cumulatieve verdeling

$$F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{+\infty} f_Z(u,y)\,dudy \tag{3.30}$$

waar we $F_X(x)$ de marginale verdeling van x noemen. We zien duidelijk dat de twee-dimensionale verdeling $f_Z(x,y)$ ondubbelzinnig de twee marginale verdelingen $F_X(x)$ en $F_Y(y)$ definiëren, maar niet omgekeerd. De twee marginale waarschijnlijkheidsdichtheidsverdelingen $F_X(x)$ en $F_Y(y)$ definiëren enkel de twee-dimensionale verdeling $f_Z(x,y)$ indien de twee stochastieken X en Y onafhankelijk zijn. Dan geldt

$$f_Z(x,y) = f_X(x)\cdot f_Y(y) \ . \tag{3.31}$$

Indien deze vergelijking niet geldt, spreken we over gecorreleerde stochastieken. In het voorbeeld over sterren zullen bijvoorbeeld de temperatuur en de druk in het centrum van de ster gecorreleerd zijn, daar er een fysisch verband bestaat. Het andere voorbeeld over dobbelstenen heeft te maken met twee ongecorreleerde of onafhankelijke variabelen, de uitkomst van de ene dobbelsteen heeft helemaal geen invloed op de uitkomst van de andere dobbelsteen.

De correlatie tussen twee stochastische variabelen X en Y kunnen we ook wiskundig kwantificeren met behulp van hun verwachtingswaarden en varianties (die we ook kunnen uitdrukken als verwachtingswaarden van momenten van de verdeling). De theoretische verwachtingswaarde [5], zoals gedefinieerd in 3.13, kunnen we gemakkelijk veralgemenen naar meer-dimensionale kansvectoren, meer in het bijzonder voor een twee-dimensionale kansvector hebben we

$$E[g(X,Y)] = \int\int_{\Omega} g(x,y)\ f_Z(x,y)\ dxdy \tag{3.32}$$

waar Ω terug de steekproefruimte voorstelt. Bijgevolg kunnen we eenvoudig de gemiddelde waarde van X en Y bepalen als

$$\mu_x = E[X] = \int\int_{\Omega} x\ f_Z(x,y)\ dxdy = \int_{-\infty}^{+\infty} x \int_{-\infty}^{+\infty} f_Z(x,y)\ dydx \tag{3.33}$$

en

$$\mu_y = E[Y] = \int\int_{\Omega} y\ f_Z(x,y)\ dxdy = \int_{-\infty}^{+\infty} y \int_{-\infty}^{+\infty} f_Z(x,y)\ dxdy \tag{3.34}$$

waar we direct gebruik kunnen maken van de marginale verdelingen. Het tweede centrale moment of de variantie bepalen we op een analoge manier

$$\mathrm{Var}[X] = \sigma_X^2 = E[(X-\mu_x)^2] = \int\int_{\Omega} (x-\mu_x)^2\ f_Z(x,y)\ dxdy \tag{3.35}$$

en

$$\mathrm{Var}[Y] = \sigma_Y^2 = E[(Y-\mu_y)^2] = \int\int_{\Omega} (y-\mu_y)^2\ f_Z(x,y)\ dxdy\ . \tag{3.36}$$

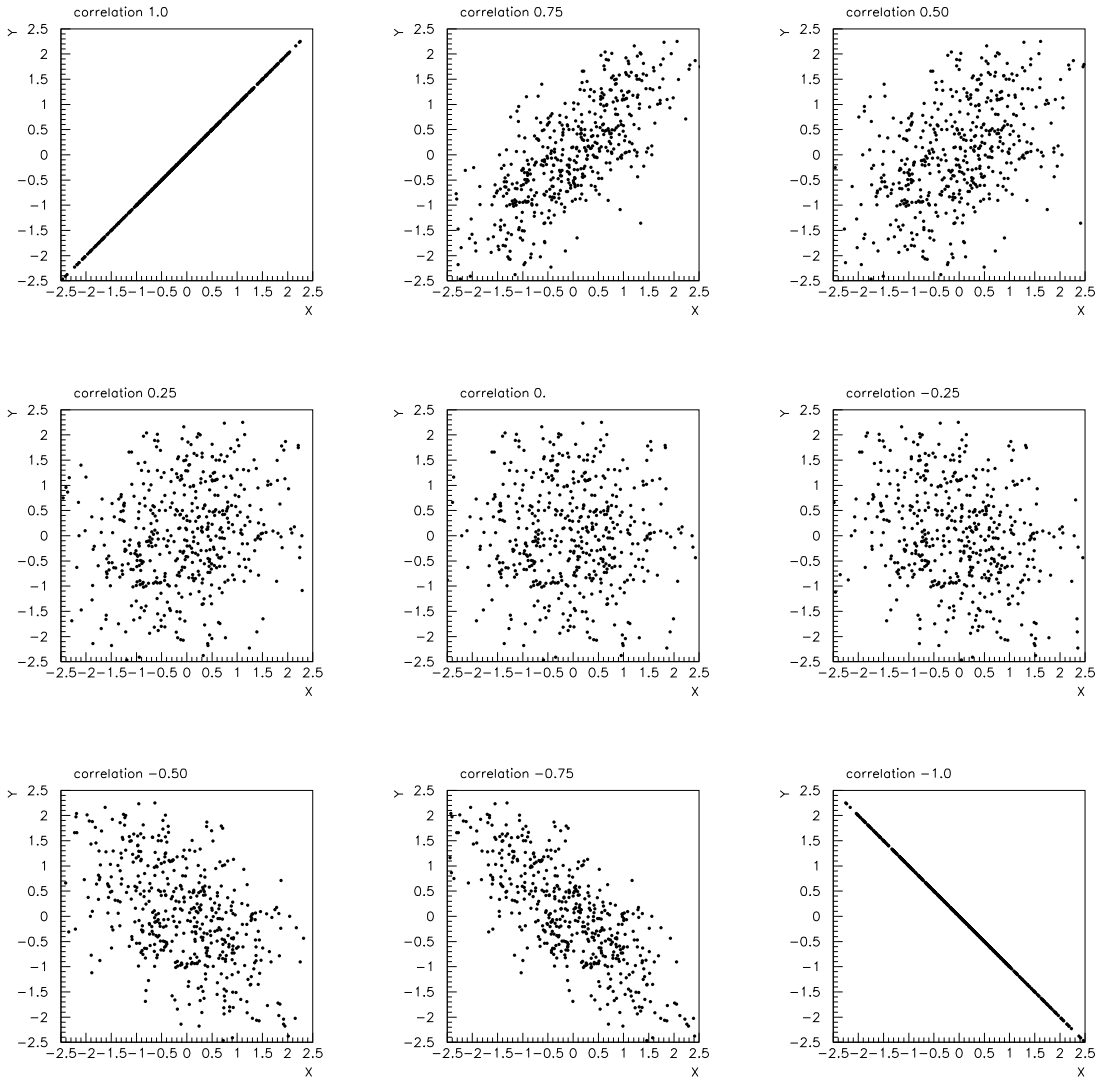
Met deze grootheden kunnen we een belangrijke grootheid definiëren die de twee-dimensionale verdeling $f_Z(x,y)$ karakteriseert, namelijk de covariantie $cov(X,Y)$ (soms ook genoteerd als σ_{xy}^2) of het gemengde centrale moment

$$\sigma_{xy}^2 = cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E[XY] - E[X]\ E[Y] \tag{3.37}$$

en bijhorende correlatiecoëfficiënt $\rho(X,Y)$

$$\rho = \frac{cov(X,Y)}{\sigma_x\ \sigma_y}\ . \tag{3.38}$$

[5] Vergeet niet dat we dit het rekenkundig gemiddelde noemen voor empirische gegevens en berekenen via sommaties over alle gegevens.



Figuur 3.3: *Illustratie van steekproeven van stochastieken X en Y voor verschillende correlatieco¨effi ci¨enten ρ.*

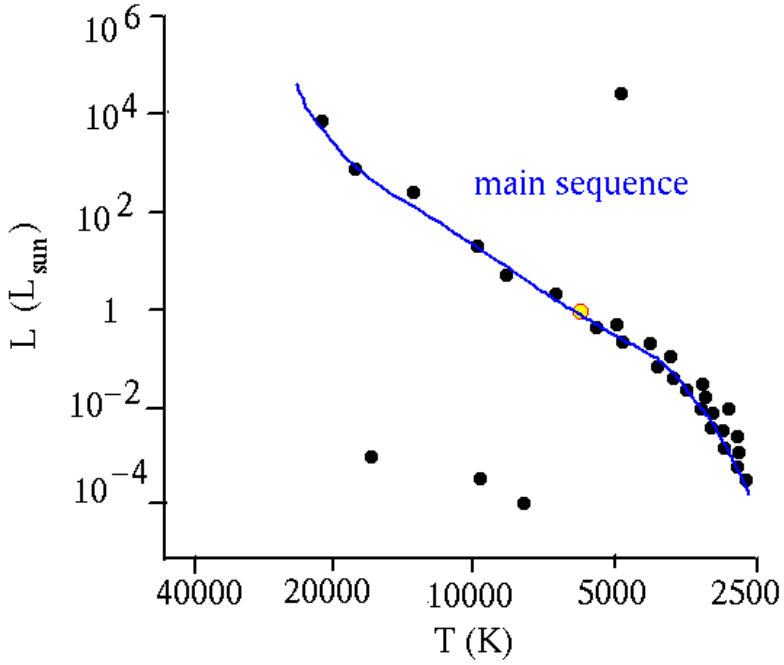
De correlatiecoëfficiënt $\rho(X, Y)$ heeft steeds een waarde die tussen -1 en 1 ligt [6]. In Figuur 3.3 vinden we een steekproef van 500 experimenten van de twee-dimensionale verdeling $f_Z(x, y)$, en dit voor verschillende correlatiecoëfficiënten. Het is duidelijk dat wanneer de absolute waarde van de correlatiecoëfficiënt $|\rho|$ groot is, de verdeling een opgelijnde structuur heeft en de stochastische variabelen een grotere afhankelijkheid hebben van elkaar, in tegenstelling tot $\rho \sim 0$ waar de lijnstructuur volledig weg is en de variabelen bijgevolg eerder onafhankelijk zijn.

Dit begrip van correlatie tussen twee (of meerdere) stochastische variabelen, eigen aan

[6] Dit wordt bewezen in de cursus 'Lineaire Algebra' en houdt verband met de ongelijkheid van Cauchy-Schwartz.

elementen uit de steekproefruimte is essentieel in de wetenschap. De Russische chemicus Dmitriy Mendeleev (1834-1907) kon via de correlaties die hij vond tussen de eigenschappen van atomen en hun atoommassa zijn bekende tabel opstellen. Bij het opstellen van nieuwe theorieën of bij het onderzoeken van nog ongekende fenomenen, gaat men meestal zoeken naar verbanden tussen waarnemingen en/of verbanden tussen grootheden. Door het kwantificeren en nadien bestuderen van deze correlaties, ontdekt men vaak nieuwe ideëen over het onderwerp.

Een voorbeeld is het zogenaamde Hertzsprung-Russell (H-R) diagram in de sterrenkunde. Een Hertzsprung-Russell diagram is een twee-dimensionale weergave van een verzameling sterren. In de Figuur 3.4 wordt elke ster weergegeven door een punt met twee bijhorende stochastische variabelen, namelijk de luminositeit van de ster (bijvoorbeeld relatief ten opzichte van de luminositeit van de ster in het midden van ons zonnenstelsel, de Zon) en de oppervlakte temperatuur van dezelfde ster. Aan de hand van dergelijke observaties kan men modellen opstellen voor de evolutie van sterren.



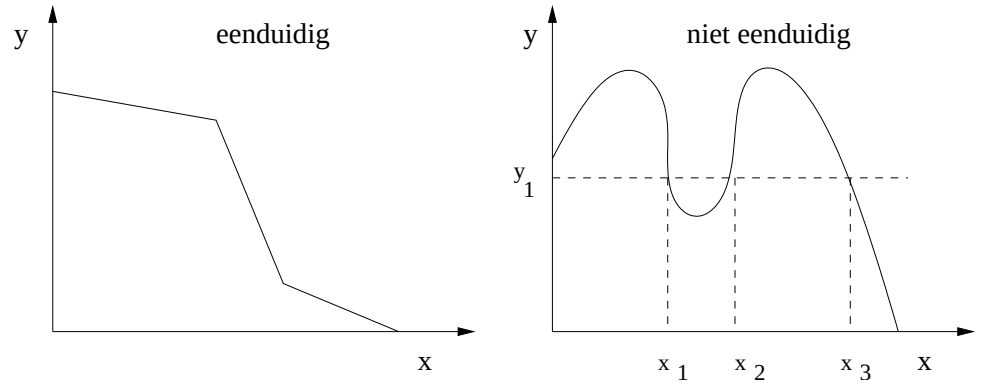
Figuur 3.4: *Illustratie van een Hertzsprung-Russell diagram als voorbeeld van een geobserveerde correlatie tussen twee stochastische variabelen van een ster.*

3.4 Bewerkingen met stochastische variabelen

In sommige toepassingen willen we uit de verdeling f_X van een stochastische variabele X, de verdeling f_Y halen van een stochastische variabele Y die gerelateerd is met X als

$$y = g(x) \tag{3.39}$$

waar $g(x)$ een functie is van x die y ondubbelzinnig bepaalt. Met ondubbelzinnig bedoelen we een functie waar de waarde van $g(x)$ eenduidig bepaald is. Dit wordt geïllustreerd in Figuur 3.5.



Figuur 3.5: *Illustratie van functies die eenduidig (one-to-one mapping) en niet eenduidig (many-to-one mapping) zijn.*

Voor een continue verdeling met cumulatieve F_X kunnen we schrijven dat

$$F_X(t) = P(X \leq t) = P(Y = g(X) \leq g(t) = u) = F_Y(u) \tag{3.40}$$

indien $g(x)$ een stijgende functie is ($dg(x)/dx \geq 0$) en

$$F_X(t) = P(X \leq t) = P(Y = g(X) \geq g(t) = u) = 1 - F_Y(u) \tag{3.41}$$

indien $g(x)$ een dalende functie is ($dg(x)/dx < 0$). Bijgevolg kunnen we stellen dat de afgeleide van de cumulatieve functie F_X

$$dF_X(x) = f_X(x)\, dx = f_X(x) \left|\frac{dx}{dy}\right| dy \tag{3.42}$$

gelijk moet zijn aan de afgeleide van de cumulatieve functie F_Y en dus geldt

$$f_Y(y) = \frac{1}{\left|\frac{dg(x)}{dx}\right|} f_X(x) \quad . \tag{3.43}$$

Indien de functie $g(x)$ niet eenduidig is, moet men sommeren over alle mogelijkheden

$$f_Y(y) = \sum_{i=1}^{n} \left(\left| \frac{dg(x)}{dx} \right|_{x=x_i} \right)^{-1} f_X(x_i) \tag{3.44}$$

waar n het aantal mogelijkheden is.

Als voorbeeld kunnen we teruggrijpen naar de verdeling van Maxwell-Boltzmann die we ook kunnen schrijven als de verdeling van de snelheid v van een molecule met massa m in een gas met absolute temperatuur T:

$$f_V(v) = \alpha v^2 e^{-\beta v^2} \quad \forall v : 0 \leq v < \infty \tag{3.45}$$

waar V de stochastische variabele is met waarden v, α is een normalisatiefactor en $\beta = m/2kT$ (k de constante van Boltzmann). Met deze kennis willen we nu weten wat de waarschijnlijkheidsdichtheidsverdeling is voor de kinetische energie E

$$E = \frac{1}{2}mv^2 \ . \tag{3.46}$$

Door deze vergelijking af te leiden bekomen we

$$dE = mvdv \tag{3.47}$$

en kunnen we bijgevolg schrijven dat ($e \to t$)

$$f_E(t) = \alpha \, \frac{2t}{m} \, e^{-\beta \frac{2t}{m}} \, \frac{1}{m} \sqrt{\frac{m}{2t}} \tag{3.48}$$

of

$$f_E(t) = \alpha' \sqrt{t} \, e^{-\beta' t} \tag{3.49}$$

met

$$\alpha' = \alpha \sqrt{\frac{2}{m^3}} \qquad \text{en} \qquad \beta' = \frac{2\beta}{m} \ . \tag{3.50}$$

Hier is E de stochatische variabele die waarden t kan aannemen (hier heb ik niet gekozen voor e als waarde voor stochastiek E om de verwaring met de exponentiaal te vermijden).

Het is soms nuttig om het empirisch gemiddelde van de som, het product of een functie van stochastische variabelen te bepalen. Voor de som en het product beschouwen we twee stochastische variabelen X en Y die waarden x en y kunnen aannemen en hebben we n metingen gedaan, namelijk $\{(x_i, y_i) \mid i \in \{1, 2, ..., n\}\}$. Voor de som $X + Y$ schrijven we het rekenkundig gemiddelde als

$$\overline{x+y} \ = \ \frac{1}{n} \sum_{i=1}^{n} (x_i + y_i) \ = \ \frac{1}{n} \sum_{i=1}^{n} x_i \ + \ \frac{1}{n} \sum_{i=1}^{n} y_i \ = \ \overline{x} + \overline{y} \tag{3.51}$$

en de variantie als

$$\sigma^2_{x+y} \ = \ \frac{1}{n-1} \sum_{i=1}^{n} \left((x_i + y_i) - (\overline{x+y}) \right)^2 \ = \ \sigma^2_x \ + \ \sigma^2_y \ + 2 \cdot cov(X, Y) \tag{3.52}$$

en indien X en Y niet gecorreleerd of onafhankelijk zijn, vereenvoudigt dit zich tot

$$\sigma^2_{x+y} \ = \ \sigma^2_x \ + \ \sigma^2_y \ . \tag{3.53}$$

Deze uitdrukkingen voor de som zijn eenvoudig om te zetten naar gelijkaardige uitdrukkingen voor het verschil van twee stochastische variabelen.

Voor het product van twee stochastische variabelen X en Y die n keer gemeten zijn, vinden we volgende relatie

$$n \cdot cov(X,Y) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i(y_i - \overline{y}) - \overline{x}\sum_{i=1}^{n}(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \quad (3.54)$$

en bijgevolg

$$\overline{x \cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \overline{x}\,\overline{y} + cov(X,Y) \ . \quad (3.55)$$

Voor de variantie van het produkt kunnen we volgende identiteit gebruiken

$$x_i y_i - \overline{x}\,\overline{y} = \overline{x}(y_i - \overline{y}) + \overline{y}(x_i - \overline{x}) + (x_i - \overline{x})(y_i - \overline{y}) \quad (3.56)$$

en de veronderstelling dat de varianties van X en Y klein zijn ten opzichte van de rekenkundige gemiddelden $\overline{x}$ en $\overline{y}$, met andere woorden dat de laatste term in het rechterlid van uitdrukking 3.56 verwaarloosbaar is ten opzichte van de twee vorige. Dit is zeker waar indien de variabelen onafhankelijk zijn. We bekomen dan voor de variantie

$$\sigma_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})^2 \simeq \frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{x}(y_i - \overline{y}) + \overline{y}(x_i - \overline{x})\right)^2 = \overline{x}^2\sigma_y^2 + \overline{y}^2\sigma_x^2 \ . \quad (3.57)$$

Het is soms nuttig om de stochastische variabele X te transformeren via een functie $g(x)$ en om van deze functie het gemiddelde $\overline{f(x)}$ en de variantie $\sigma_{f(x)}^2$ te bepalen. Om niet voor iedere empirische meetwaarde afzonderlijk de functiewaarde $f(x_i)$ te hoeven uitrekenen, willen we de relatie nagaan tussen $\overline{f(x)}$ en $f(\overline{x})$. Dit kunnen we doen door bijvoorbeeld het verschil tussen beide te bepalen

$$\overline{f(x)} - f(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - f(\overline{x})\right) \ . \quad (3.58)$$

Indien de functie $f(x)$ tweemaal afleidbaar is, kunnen we de functie benaderen met een Taylorontwikkeling rond $x_i = \overline{x}$

$$f(x_i) = \sum_{n=0}^{\infty}\frac{f^{(n)}(\overline{x})}{n!}(x_i - \overline{x})^n \ = \ f(\overline{x}) + (x_i - \overline{x})f'(\overline{x}) + \frac{1}{2}(x_i - \overline{x})^2 f''(\overline{x}) + \ldots \quad (3.59)$$

waar we alle termen van hogere orde ($n \geq 3$) verwaarlozen. We verkrijgen bijgevolg

$$\overline{f(x)} - f(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - f(\overline{x})\right) \simeq \frac{1}{n}\sum_{i=1}^{n} f'(\overline{x})(x_i - \overline{x}) + \frac{1}{2n}\sum_{i=1}^{n} f''(\overline{x})(x_i - \overline{x})^2 \quad (3.60)$$

waar de eerste term in het rechterlid gelijk is aan nul. Door gebruik te maken van de vergelijking

$$f(x_i) - \overline{f(x)} = f(x_i) - f(\overline{x}) + f(\overline{x}) - \overline{f(x)} \tag{3.61}$$

kunnen we volgende benadering maken voor de variantie $\sigma^2_{f(x)}$

$$\sigma^2_{f(x)} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(f(x_i) - \overline{f(x)}\right)^2 \tag{3.62}$$

of door gebruik van vergelijking 3.61

$$\sigma^2_{f(x)} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(f(x_i) - f(\overline{x})\right)^2 \;-\; \frac{n}{n-1}\left(\overline{f(x)} - f(\overline{x})\right)^2 \tag{3.63}$$

en via de Taylorontwikkeling 3.59

$$\sigma^2_{f(x)} \;\simeq\; \frac{1}{n-1}\sum_{i=1}^{n}\left(f'(\overline{x})(x_i - \overline{x}) + \frac{1}{2}f''(\overline{x})(x_i - \overline{x})^2\right)^2 - \frac{n}{n-1}\left(\overline{f(x)} - f(\overline{x})\right)^2 \tag{3.64}$$

en als we alles uitschrijven en gebruik maken van het verschil 3.60 vinden we

$$\sigma^2_{f(x)} \;\simeq\; f'(\overline{x})^2\,\sigma_x^2 \;. \tag{3.65}$$

Dit geldt indien $|(x_i - \overline{x})f''(\overline{x})|$ veel kleiner is dan $|f'(\overline{x})|$ voor alle $i \in \{1, 2, ..., n\}$. Indien deze tweede afgeleide inderdaad klein is, kunnen we ook stellen dat het verschil 3.60 klein is en dus

$$\overline{f(x)} \simeq f(\overline{x}) \tag{3.66}$$

of het rekenkundig gemiddelde van de empirische gegevens $f(x)$ ongeveer gelijk is aan de functiewaarde van het rekenkundig gemiddelde $\overline{x}$. De benaderingen 3.65 en 3.66 zijn exact indien de functie $f(x)$ lineair is, de tweede afgeleide is dan namelijk gelijk aan nul.

Analoog willen we de verwachtingswaarde $E[Z]$ kennen van de som $Z = X + Y$, het product $Z = XY$ of een functie $Z = f(X)$ van stochastische variabelen die voorgesteld worden door een theoretische verdeling in plaats van aan de hand van empirische gegevens. Via de definitie van de verwachtingswaarde 3.32 kunnen we eenvoudig volgende uitdrukkingen bewijzen:

- $E[X + Y]$ = $E[X] + E[Y]$;
- $E[XY]$ = $E[X]\ E[Y]$;
- Var$[X + Y]$ = Var$[X]$ + Var$[Y]$ + 2 $\rho(X, Y)\ \sqrt{\text{Var}[X]\text{Var}[Y]}$;
- Var$[XY]$ = Var$[X]$ Var$[Y]$ + $E[X]^2$ Var$[Y]$ + $E[Y]^2$ Var$[X]$.

Waar de eigenschappen voor het product XY enkel gelden indien de stochastische variabelen X en Y onafhankelijk zijn. Deze eigenschappen kan men uitschrijven voor zowel discrete als continue verdelingen. Voor onafhankelijke of ongecorreleerde stochastische variabelen bekomen we welgekende uitdrukking

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] \tag{3.67}$$

die de variantie van de som (of verschil) van twee variabelen gelijkstelt aan de som van de varianties van de twee afzonderlijke variabelen. We zullen later zien hoe we de variantie van een functie $Z = g(\{X_i\})$ bepalen aan de hand van de varianties van de verzameling individuele stochastische variabelen $\{X_i\}$ en hun onderling functieverband.

3.5 Grafische voorstelling van empirische gegevens

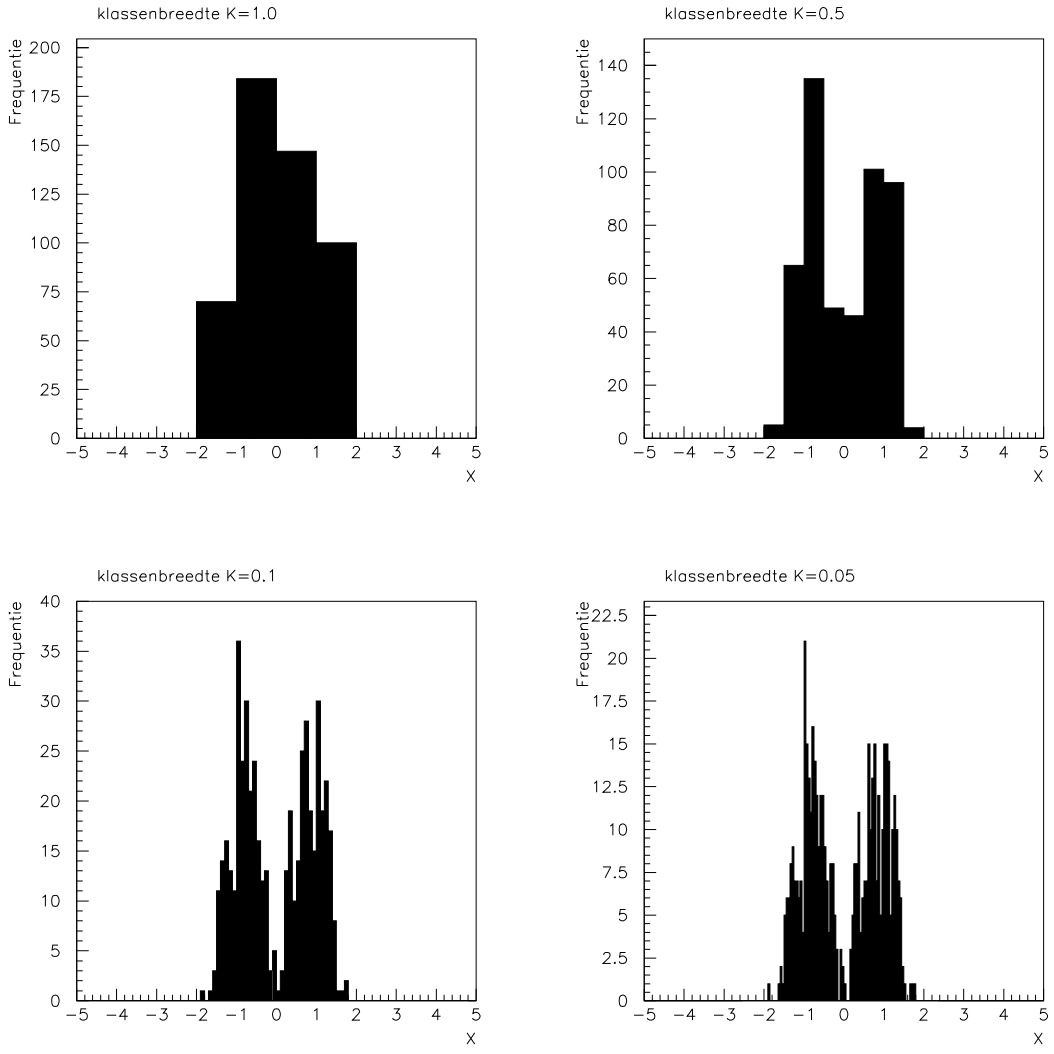
De theoretische verdelingen kunnen we voorstellen door hun analytische functie. Empirische gegevens bekomen uit een eindige set metingen echter, moet men voorstellen aan de hand van ruwe data $\{x_i \mid i \in \{1,2,...,n\}\}$. Deze ruwe data gebruiken we bijvoorbeeld om het gemiddelde $\overline{x}$, de standaardafwijking s_x of de verschillende momenten μ_k of μ'_k van de verzameling gegevens te bepalen. Maar om deze data visueel voor te stellen in een publicatie, moeten we de ruwe data compacter neerschrijven.

Dit gebeurt door middel van een zogenaamd histogram. Een histogram wordt gevormd door het verzamelen van ruwe data $\{x_i\}$ in N klassen $\{k_{j-1} < x_i \leq k_j \mid j \in \{1,2,...,N\}\}$. Een parameter die het histogram definieert is de klassenbreedte $K = k_j - k_{j-1}$ die meestal constant genomen wordt voor elke waarde van j. Voor elke meting x_i gaan we na in welk half-open interval j deze geklasseerd kan worden. Indien we dit gedaan hebben voor alle n metingen, kunnen we de frequentie f_j bepalen voor elke klasse j. Als klasse $j = 3$ bijvoorbeeld bevolkt wordt door 14 van de n metingen $\{x_i\}$, danis de frequentie van deze klasse gelijk aan $f_3 = 14$. Bijgevolg geldt

$$\sum_{j=1}^{N} f_j = n \tag{3.68}$$

dat de som van alle frequenties over alle klassen gelijk moet zijn aan het totaal aantal uitgevoerde empirische metingen. Elke meting moet namelijk ergens voorkomen. In een histogram wordt bijgevolg alle ruwe data $\{x_i\}$ gesorteerd. Deze ruwe data worden nadien verdeeld in klassen van meestal dezelfde breedte. In de grafische voorstelling representeren we elke klasse j door een rechthoek met breedte gelijk aan de klassenbreedte en een hoogte die overeenkomt met de frequentie f_j van de klasse. De oppervlakte S_i van de rechthoek is bijgevolg evenredig met de frequentie f_j, namelijk $S_i = K \cdot f_j$.

In Figuur 3.6 vinden we vier voorbeelden van histogrammen die dezelfde verzameling ruwe data weergeven maar met verschillende klassenbreedten, $K = 1.0, 0.5, 0.1$ en 0.05. Het is duidelijk dat, indien de klassebreedte te groot is, men de structuur van de verdeling niet weergeeft. Het histogram met klassenbreedte $K = 1.0$ heeft namelijk slechts één piek, terwijl de ruwe data duidelijk rond twee verschillende waarden gespreid liggen, namelijk $x = -1.0$ en $x = 1.0$. Ook indien de klassenbreedte te klein wordt, bijvoorbeeld $K = 0.05$,



Figuur 3.6: *Illustratie van een ´e´en-dimensionaal histogram van 500 empirische gegevens van stochastiek X, dit voor verschillende klassenbreedten K.*

verdwijnt de algemene structuur van de verdeling. Het kiezen van de waarde van de klassenbreedte is bijgevolg essentieel en kan ook misbruikt worden om bepaalde structuren van de verdeling te verdoezelen. In de praktijk dient men het aantal klassen N te kiezen aan de hand van het aantal empirische metingen n. Hoe groter n, hoe kleiner we de klassenbreedte K kunnen nemen en bijgevolg hoe meer klassen N er zijn.

Een histogram kan helpen bij het oplossen van volgende vragen:

- Wat is de onderliggende theoretische verdeling van de ruwe data ?
- Wat is de ligging van de ruwe data ?
- Hoe groot is de spreiding van de ruwe data ?

- Volgt de data al dan niet een symmetrische verdeling ?
- Zijn er uitschieters die misschien op meetfouten duiden ?

Het is ook eenvoudig om de waarschijnlijkheidsdichtheidsverdeling af te leiden uit het absolute histogram. Men kan namelijk alle frequenties f_j delen door het totaal aantal empirische metingen n

$$P_j = \frac{f_j}{n} \tag{3.69}$$

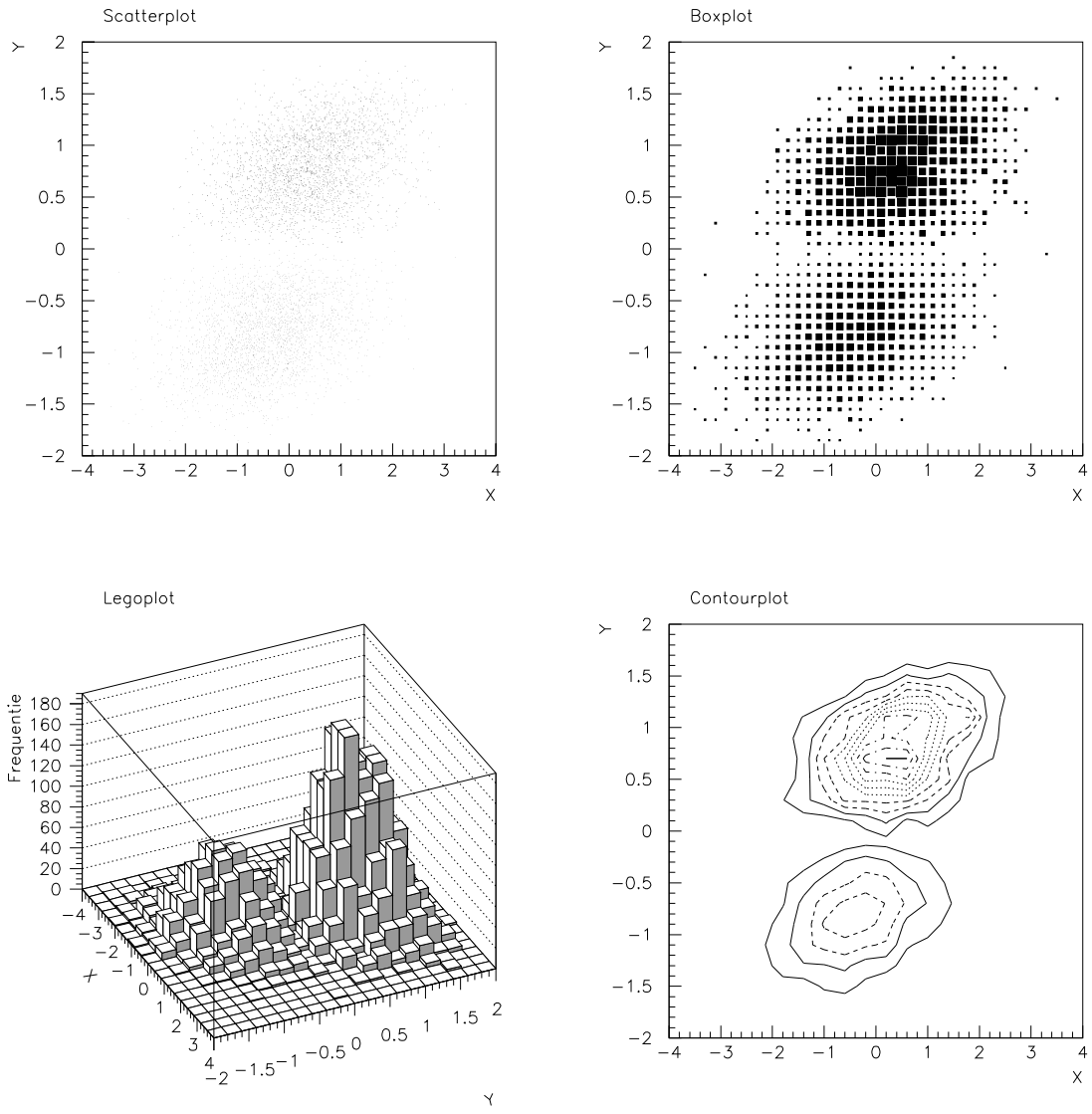
waar P_j de empirische waarschijnlijkheid voorstelt van iedere klasse j.

Twee nieuwe grootheden kunnen gedefinieerd worden aan de hand van het histogram, namelijk de modus en de modale klasse. De modus van een steekproef is de meest voorkomende waarneming. Voor een discrete stochastiek is dit meestal een nuttige grootheid, daar indien het aantal metingen n groter is dan het aantal elementen in de steekproefruimte Ω men steeds elementen $x_i \in \Omega$ heeft die meerdere keren voorkomen. In het geval van een continue stochastiek X die alle waarden $x \in \Re = \Omega$ kan aannemen, komt elke waarde meestal slechts één keer voor. Na het catalogeren van de ruwe gegevens in een histogram, kan men wel spreken over de klasse met de grootste frequentie f_j. Men noemt deze klasse j de modale klasse.

Het concept van een één-dimensionaal histogram kunnen we uitbreiden naar meerdere dimensies d. Hiervoor moeten we enkel d-dimensionale klassen definiëren die bevolkt worden met de empirische metingen van de stochastieken $(X_1, X_2, ..., X_d)$ toegekend aan de n elementen. Het totaal aantal metingen die binnen klasse $(j_1, j_2, ..., j_d)$ vallen geeft de frequentie $f_{(j_1,j_2,...,j_d)}$ weer. In twee dimensies is dergelijk histogram nog eenvoudig voor te stellen. Voor meerdere dimensies kunnen we de marginale verdelingen van de d-dimensionale verdeling voorstellen aan de hand van een histogram.

In Figuur 3.7 vinden we vier voorbeelden hoe we een twee-dimensionale verdeling van stochastieken X en Y grafisch kunnen voorstellen.

- De **scatterplot**: Elke meting van de steekproef wordt voorgesteld door een punt in het twee-dimensionaal vlak (X, Y), waar de waarden van de coördinaten overeenkomen met de empirisch gemeten waarde van de stochastieken. Op deze manier kan men een overzicht krijgen van waar de data zich bevindt en kan men nog naar individuele uitschieters zoeken.
- Het **twee-dimensionaal histogram** of **legoplot**: Dit is de twee-dimensionale voorstelling van de verzameling gegevens, opgesplitst in klassen. De hoogte van de balk geeft de frequentie van de klasse weer.
- De **boxplot**: Dit is een gelijkaardige voorstelling als het histogram, men moet namelijk ook de data ordenen in klassen. Hier is de oppervlakte van het vierkant evenredig met de frequentie van de klasse. Deze voorstelling is soms duidelijker dan het eigenlijke histogram zelf.



Figuur 3.7: *Illustraties van eenzelfde twee-dimensionale verdeling van 5000 empirische gegevens van stochastiek X en Y.*

- De **contourplot**: In deze voorstelling worden in het twee-dimensionaal vlak (X, Y) punten (effectieve middens van klassen) verbonden met elkaar, indien ze dezelfde frequentie hebben. Men kan spreken van iso-frequente lijnen, naar analogie met isobare of isotherme lijnen die men terugvindt op weerkaarten. Meestal maakt men gebruik van een interpolatie tussen middens van klassen, daar men zelden twee klassen vindt die dezelfde frequentie hebben.

We zien twee gescheiden ophopingen van empirische gegevens rond de (x, y)-coördinaten $(+1.0, +1.0)$ en $(-1.0, -1.0)$. In de scatterplot is het moeilijk te zien dat de ene ophoping meer bevolkt is dan de andere, terwijl dit in de andere voorstellingen duidelijk zichtbaar is.

De correlatie tussen de stochastieken X en Y is aan de andere kant bijna niet zichtbaar indien men het histogram of de legoplot bekijkt, en wel duidelijk zichtbaar indien met de andere voorstellingen bekijkt. De boxplot en de contourplot geven bijna geen informatie over het aantal empirische gegevens, terwijl men dit eenvoudigweg kan aflezen in het histogram of de legoplot. Afhankelijk van wat men wil tot uiting brengen, zal men een keuze moeten maken in verband met hoe men de empirische gegevens grafisch voorstelt.

Indien men de twee-dimensionale verdeling wil normaliseren, zal enkel het histogram of de legoplot veranderen, de box- en contourvoorstellingen veranderen niet.

Hoofdstuk 4

Standaard verdelingen

"Mathematics is the queen of sciences and arithmetics the queen of mathematics. She often condescends to render service to astronomy and other natural sciences, but in all relations she is entitled to the first rank''

C.F. Gauss,
Sartorius von Walterhausen: Gauss zum Gedächtniss, 1856

Vele fysische processen hebben een stochastisch karakter en volgen een theoretische kansverdeling. Bijgevolg is het interessant om de onderliggende wiskundige verdelingen, die dergelijke fysische processen beschrijven, te bestuderen. Veelal zal men tijdens het bestuderen van empirische gegevens hun waarschijnlijkheidsdichtheidsverdeling trachten te benaderen door de verwachtte onderliggende theoretische verdeling. In dit hoofdstuk beschrijven we de belangrijkste verdelingen, alsook hun eigenschappen. De eerste twee voorbeelden zijn discrete verdelingen, terwijl de andere continue verdelingen zijn. Waar nodig wordt er verwezen naar een typische toepassing uit de wereld van de Natuurkunde.

4.1 De binomiaal verdeling

In Hoofdstuk 3 hebben we een Bernoulli experiment gedefinieerd als een stochastiek X met een steekproefruimte Ω die bestaat uit slechts twee elementen a en b. We hebben dus een discrete stochastische variabele die een waarschijnlijkheid $0 \leq p \leq 1$ heeft voor uitkomst a en een waarschijnlijkheid $q = 1 - p$ heeft voor uitkomst b. We zeggen bijgevolg dat X een Bernoulli verdeling beschrijft en noteren dit als

$$X \sim B(1, p) \tag{4.1}$$

en bekomen eenvoudige uitdrukkingen voor de verwachtingswaarde (a$\rightarrow$1 en b$\rightarrow$0)

$$E[X] = p \tag{4.2}$$

en de variantie

$$\mathrm{Var}[X] = p(1-p) \ . \tag{4.3}$$

Door het achtereenvolgens toepassen van verschillende Bernoulli experimenten, hebben we de binomiaalwet afgeleid (zie bijvoorbeeld het Bord van Galton). Als we n onafhankelijke Bernoulli experimenten (met uitkomsten 0 of 1) met kans p uitvoeren en de uitkomsten optellen, krijgen we een som Y van n Bernoulli stochastieken $X_1, X_2, ..., X_n$ die allen verdeeld zijn volgens dezelfde theoretische verdeling $B(1,p)$. De stochastiek Y geeft het aantal successen in n pogingen en neemt dus gehele waarden aan tussen 0 en n. We zeggen dat de stochastiek Y

$$Y = X_1 + X_2 + ... + X_n \tag{4.4}$$

binomiaal verdeeld is en noteren dit met

$$Y \sim B(n,p) \ . \tag{4.5}$$

De discrete waarschijnlijkheidsdichtheidsverdeling van deze stochastiek is bijgevolg

$$f_Y(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \qquad \forall k \in \{1, 2, ..., n\} \tag{4.6}$$

de verwachtingswaarde wordt

$$E[Y] = E[X_1] + E[X_2] + ... + E[X_n] = np \tag{4.7}$$

en de variantie

$$\mathrm{Var}[Y] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + ... + \mathrm{Var}[X_3] = np(1-p) \ . \tag{4.8}$$

In Figuur 2.3 vinden we enkele voorbeelden van binomiale verdelingen.

Een toepassing van de binomiaalverdeling vinden we bij het construeren van deeltjesdetectoren. We willen namelijk de sporen meten van geladen deeltjes, afkomstig van de kosmische straling en dit met behulp van dradenkamers die een efficiëntie hebben van 90%. Dit komt erop neer dat wanneer een deeltje door de dradenkamer passeert er een waarschijnlijkheid p van 90% is dat er een electrisch signaal geregistreerd wordt en met andere woorden het deeltje gelocaliseerd kan worden. Om de baan van een deeltje te kunnen bepalen, hebben we minstens drie meetpunten nodig. Bijgevolg hebben we minstens drie lagen van dradenkamers nodig. Maar we willen hierbovenop de kostprijs van het experiment zo laag mogelijk houden en toch een waarschijnlijkheid van 99% hebben dat we de geladen deeltjes detecteren en een spoor kunnen opmeten. Hoeveel dradenkamers, N, hebben we minstens nodig om die detectie waarschijnlijkheid van 99% te behouden? Dit is een typisch binomiaal probleem, daar er slechts twee uitkomsten zijn : 'detecteren' of 'niet detecteren' en het detecteren gebeurt met een waarschijnlijkheid p van 90%. Stel $P(i,p,N)$ gelijk aan de waarschijnlijkheid dat er i successen zijn uit N, als men weet dat de kans op succes voor het individueel Bernoulli experiment (één dradenkamer) gelijk is aan p. Voor slechts drie dradenkamers is de waarschijnlijkheid dat er in alle drie een detectie plaatsvindt, gelijk aan

$P(3, 0.90, 3) = 0.90^3 = 0.729$ of ongeveer 73%. Voor N dradenkamers is de waarschijnlijkheid dat minstens drie dradenkamers detecteren gelijk aan

$$P(3, 0.90, N) = \sum_{i=3}^{N} P(i, 0.90, N) = \sum_{i=3}^{N} 0.90^i (1 - 0.90)^{N-i} \begin{pmatrix} N \\ i \end{pmatrix} \tag{4.9}$$

wat we kunnen oplossen naar N met de voorwaarde dat $P(3, 0.90, N) \geq 0.99$. Voor $N = 5$ vinden we een waarschijnlijkheid van ongeveer 99.1% dat het deeltje dat door de volledige detector passeert, zal gedetecteerd worden in minstens drie lagen.

4.2 De Poisson verdeling

Het kan gebeuren dat sommige gebeurtenissen zeer zelden voorkomen, met ander woorden dat de waarschijnlijkheid p zeer klein is. Denk bijvoorbeeld (gelukkig) aan verkeersongelukken. Niettegenstaande we veel met de auto rijden, gebeurt het zelden dat we betrokken zijn in een botsing. Het is slechts omdat er veel mensen met de auto rijden en ze dat ook in totaal over lange afstanden doen, dat we iets kunnen zeggen over bijvoorbeeld de relatie tussen het dragen van een gordel en de ernst van het letsel, opgelopen tijdens een botsing. Bij dergelijke onderzoeken kan de Poisson verdeling een nuttige verdeling zijn. De Poisson verdeling was voor het eerst neergeschreven door de Fransman Siméon Denis Poisson in 1837 (zie Figuur 4.1 voor een portret) en blijkbaar voor het eerst toegepast door het Pruisische leger in 1898 om iets te leren over de al dan niet dodelijke afloop van soldaten die vertrappeld worden door hun paard.

De Poisson verdeling is een theoretische verdeling die waarschijnlijkheden toekent aan het aantal keer een verschijnsel optreedt, ze geeft de waarschijnlijkheid $P_r(t)$ dat een verschijnsel exact r keer optreedt in een gegeven tijdsinterval t in de veronderstelling dat de verschijnselen onafhankelijk van elkaar optreden.

Figuur 4.1: *Portret van Sim´eon Denis Poisson (1781-1840).*

Beschouw de stochastische variabele Y die het aantal verkeersongelukken weergeeft per dag met verwachtingswaarde $E[Y] = N$. Kunnen we dan de spreiding of variantie $\mathrm{Var}[Y]$ bepalen? Per uur wordt de verwachtingswaarde dan $N/24$ en per week $7N$, met andere woorden is de verwachtingswaarde evenredig met de lengte van de observatietijd. Stel dan τ gelijk aan een fractie van de dag, dan geeft stochastiek X_τ het aantal verkeersongelukken dat in deze fractie van de dag plaatsvindt, en geldt $E[X_\tau] = \tau N$. Nu kunnen we de observatietijd τ zo klein kiezen dat er hoogstens één verkeersongeluk plaatsvindt in die tijdsspanne en neemt X_τ slechts twee waarden aan 0 en 1. Dit komt terug neer op een Bernoulli experiment $B(1, p)$ met een waarschijnlijkheid gedefinieerd aan de hand van de

verwachtingswaarde, $p = E[X_\tau] = \tau N$ en een variantie $\mathrm{Var}[X_\tau] = \tau N(1-\tau N)$. Als Y het aantal verkeersongelukken op een dag is en een dag is opgedeeld in n identieke stukjes van lengte $\tau = 1/n$ dag, dan geldt

$$Y = \sum_{i=1}^{n} X_{\tau,i} \qquad \text{met } n\tau = 1 \text{ dag} \tag{4.10}$$

Hierbij is $X_{\tau,i}$ het aantal verkeersongelukken in het i^{de} tijdsinterval met lengte τ. De stochastische variabelen $X_{\tau,1}, X_{\tau,2}, ..., X_{\tau,n}$ zijn onafhankelijk en hebben allen dezelfde verdeling. Bijgevolg geldt:

$$E[Y] = n\ E[X_\tau] \simeq n\ \tau\ N = N \tag{4.11}$$

en

$$\mathrm{Var}[Y] = n\ \mathrm{Var}[X_\tau] \simeq n\ \tau\ N\ (1-\tau\ N) = N\ (1-\tau\ N)\ . \tag{4.12}$$

De gelijkheden worden exact indien men de limiet neemt naar heel kleine tijdsintervallen (dan kunnen zeker geen twee verschijnselen plaatsvinden in één tijdsinterval) of $\tau \to 0$ en bijgevolg ook voor $n \to \infty$. Alleen met de veronderstellingen dat de verwachtingswaarde per tijdsinterval constant is in de tijd en dat de aantallen verschijnselen in twee exclusieve tijdsintervallen onafhankelijke stochastieken zijn, kunnen we de Poisson waarschijnlijkheidsdichtheidsverdeling volledig en eenduidig definiëren.

De Poissonverdeling $P(\lambda)$ waar λ de verwachtingswaarde is, kunnen we vinden uit de binomiaal verdeling $B(n, \lambda/n)$ door de limiet voor $n \to \infty$ te nemen, dus als

$$Y \sim P(\lambda) \tag{4.13}$$

dan geldt

$$P(k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \lim_{n\to\infty} \frac{\lambda^k}{k!} \frac{n!}{n^k (n-k)!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \tag{4.14}$$

waar we enkele limieten kunnen gebruiken [1]

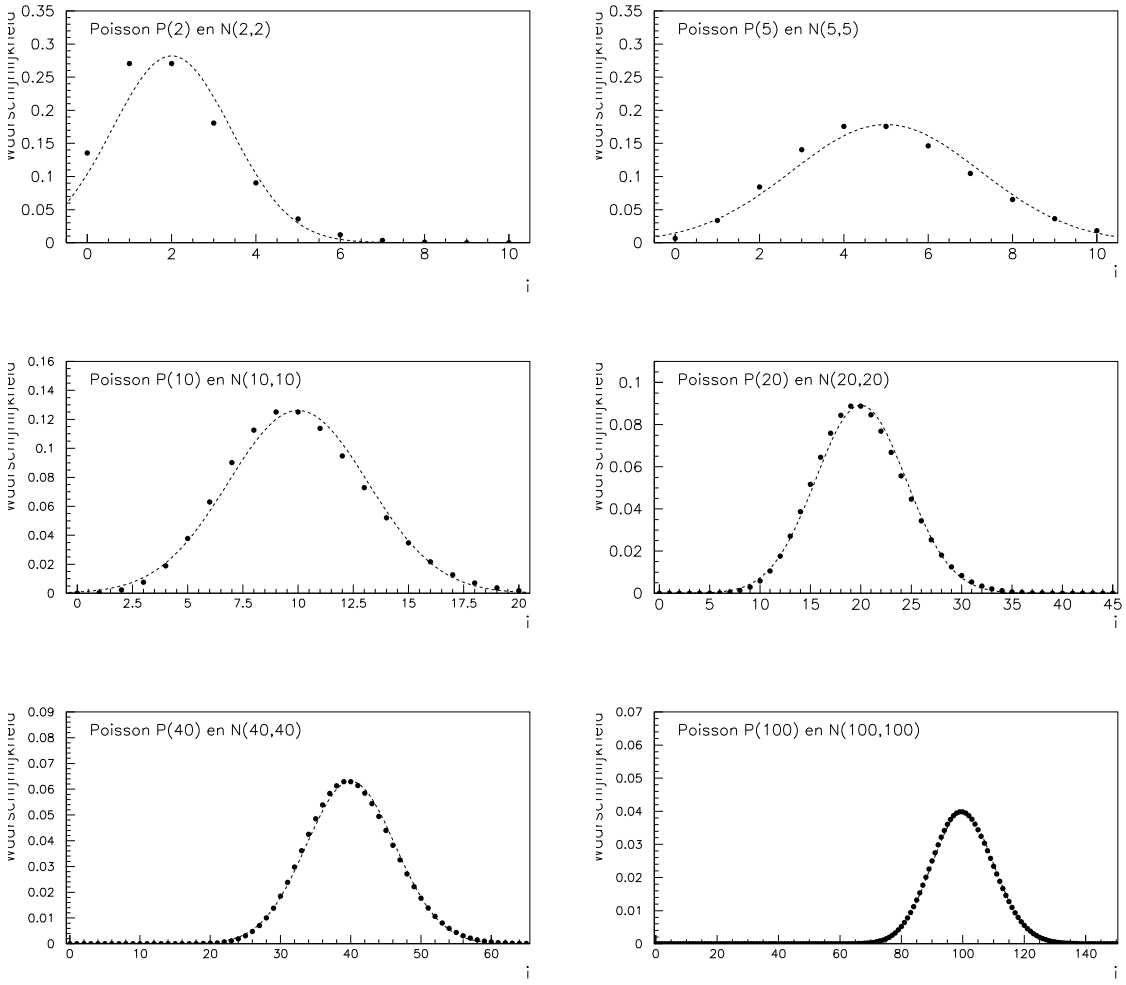
$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \tag{4.15}$$

en

$$\lim_{n\to\infty} \frac{n!}{n^k\ (n-k)!} \left(1-\frac{\lambda}{n}\right)^{-k} = 1 \tag{4.16}$$

om uiteindelijk de Poissonverdeling te bekomen

$$P_Y(k) = \frac{\lambda^k}{k!}\, e^{-\lambda}\ . \tag{4.17}$$



Figuur 4.2: *Enkele voorbeelden van de discrete Poisson verdeling met verwachtingswaarde μ. Ook de continue normale of Gaussiaanse verdeling $N(\mu, \mu)$ wordt getoond als een stippellijn.*

Het is eenvoudig na te gaan dat de totale waarschijnlijkheid inderdaad gelijk is aan 1

$$\sum_{k=0}^{\infty} P_Y(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}\, e^{\lambda} = 1 \quad . \tag{4.18}$$

De verwachtingswaarde is

$$E[Y] = \sum_{k=0}^{\infty} k\, P_Y(k) = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda}\, \lambda\, e^{\lambda} = \lambda \tag{4.19}$$

en de variantie

[1] Voor een afleiding van deze limieten moet ik verwijzen naar andere cursussen zoals Integraal- en Differentiaalrekening.

$$\mathrm{Var}[Y] = E[Y^2] - E[Y]^2 = E[Y(Y-1)] + E[Y] - E[Y]^2 = \lambda \qquad (4.20)$$

waar we gebruik hebben gemaakt van de relatie

$$E[Y(Y-1)] = \sum_{k=0}^{\infty} k(k-1)\, P_Y(k) = e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} = e^{-\lambda}\, \lambda^2 e^{\lambda} = \lambda^2 \quad . \qquad (4.21)$$

De binomiaal verdeling benadert een Poisson verdeling indien de waarschijnlijkheid p dat een verschijnsel zich voordoet zeer klein is. Maar voor de binomiaal verdeling moeten we helaas twee gegevens kennen om de verdeling te definiëren, namelijk zowel de auto's zonder verkeersongeluk en het totaal aantal auto's op de weg. Om de Poisson verdeling te specifiëren moeten we slecht het eerste kennen, wat in de praktijk meestal eenvoudiger is.

De Poisson verdeling benadert een normale of Gaussiaanse verdeling (zie verder in dit hoofdstuk) indien de verwachtingswaarde $E[Y]$ groot wordt. De normale of Gaussiaanse verdeling die een Poisson verdeling met verwachtingswaarde μ benadert heeft op zijn beurt een verwachtingswaarde gelijk aan μ en een variantie gelijk aan μ, namelijk $N(\mu, \mu)$ zoals we dit zullen noteren. Dit kunnen we illustreren aan de hand van enkele voorbeelden in Figuur 4.2.

De Poisson verdeling wordt gebruikt voor verschillende toepassingen, daar men de tijdsintervallen kan vervangen door elke continue meetbare grootheid zoals lengte, oppervlakte, volume, temperatuur, enzovoort. Waarschijnlijk het meest bekende voorbeeld is de studie van het radioactief verval van atoomkernen. De atoomkern kan namelijk 'vervallen' of 'niet vervallen' gedurende het tijdsinterval t. Omdat radioactief verval een statistisch proces is, is het onmogelijk om te voorspellen of een bepaalde atoomkern vroeg dan wel laat vervalt. Men kan op die manier de halveringstijd definiëren als de tijd binnen dewelke gemiddeld de helft van het aantal atoomkernen vervalt [2]. Zie ook verder bij de exponentiële verdeling.

Het Irvine-Michigan-Brookhaven experiment heeft op 23 februari 1987 een aantal neutrino gebeurtenissen waargenomen in tijdsintervallen van 10 seconden. Dit op het moment dat de supernova S1987a voor het eerst gezien werd door astronomen. Ze bekomen volgende resultaten:

aantal neutrino's	0	1	2	3	4	5	6	7
aantal keer opgemeten in 10 seconden	1042	860	307	78	15	3	0	0
Poisson voorspelling	1064	823	318	82	16	2	0.3	0.03

Indien we willen verifiëren of deze gegevens een Poisson verdeling volgen, moeten we het gemiddelde λ bepalen van het aantal neutrino's waargenomen gedurende een tijdsinterval van 10 seconden.

[2] Deze begrippen komen uitgebreid terug in de cursussen Algemene Natuurkunde.

$$\lambda = \frac{(\text{frequentie}) \cdot (\text{aantal neutrino's})}{\Delta T} = \frac{860 + 307 \cdot 2 + 78 \cdot 3 + 15 \cdot 4 + 3 \cdot 5}{1042 + 860 + 307 + 78 + 15 + 3} = 0.77 \tag{4.22}$$

De waarschijnlijkheden die overeenkomen met een Poisson verdeling $P(\lambda)$ met verwachtingswaarde $\lambda = 0.77$ komt goed overeen met de geregistreerde gegevens. In bovenstaande tabel vinden we deze voorspelling terug, waar we de waarschijnlijkheden normaliseren voor de totale observatietijd ΔT. De achtergrond gebeurtenissen, welke neutrino's zijn die afkomstig zijn van de Zon, worden goed beschreven door deze Poisson verdeling. Het experiment detecteerde ook gedurende 1 enkel tijdsinterval in totaal 9 neutrino gebeurtenissen. Voor een Poisson verdeling met een verwachtingswaarde λ van 0.77 neutrino's per 10 seconden geeft voor dergelijke gebeurtenis slechts een verwachting van ongeveer 0.0003 gebeurtenissen in een totaal tijdsinterval ΔT. Deze gebeurtenissen heeft bijgevolg een zéér kleine kans om achtergrond van de Zon te zijn, en kan men associëren met het signaal van de supernova.

4.3 De uniforme verdeling

Deze eenvoudige theoretische waarschijnlijkheidsdichtheidsverdeling f_X beschrijft een stochastiek X die een identieke waarschijnlijkheid geeft aan alle waarden x (discreet of continue). Het functievoorschrift bekomt men door rekening te houden met de voorwaarde dat de totale waarschijnlijkheid gelijk moet zijn aan 1. Binnen het interval $[a, b]$ hebben we bijgevolg een verdeling [3]

$$f_X(x) = \frac{1}{b-a} \qquad \forall x \in [a, b] \ . \tag{4.23}$$

De verwachtingswaarde bekomen we door integratie

$$E[X] = \int_a^b \frac{1}{b-a}\, x\, dx = \frac{a+b}{2} \tag{4.24}$$

wat de oppervlakte van de rechthoek is. Voor de variantie bekomen we

$$\text{Var}[X] = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\, dx = \frac{(b-a)^2}{12} \ . \tag{4.25}$$

De uniforme distributie wordt veel gebruikt om random getallen te genereren. Indien men een statistisch proces wil simuleren met de computer gaat men meestal uit van random getallen die men kan genereren als een uniforme verdeling binnen het interval $[0, 1]$. Via zogenaamde Monte Carlo technieken kan men deze uniforme verdeling omzetten in een willekeurige verdeling die bijvoorbeeld het statistisch proces onder studie benadert.

Een voorbeeld hiervan is de som van twee random getallen. Als men op een student wacht waarmee je een afspraak hebt tussen tijdstip a en $a + 10$ minuten om enkele begrippen uit de cursus te herhalen, kunnen we stochastiek X definiëren als de tijd die we moeten wachten. Die tijd is bijgevolg uniform verdeeld tussen 0 en 10 minuten. Stel dat je nadien

[3] Eenvoudig om te zetten naar een discrete stochastiek.

hebt afgesproken met een andere student om een ander deel van de leerstof te verduidelijken. Ook voor deze heb je in een tijdsinterval van 10 minuten afgesproken. We kunnen bijgevolg een tweede stochastiek Y definiëren als de tijd die je op de tweede student moet wachten. Omdat de professor ook andere bezigheden heeft dan wachten, besluit hij 's morgens de waarschijnlijkheidsdichtheidsverdeling te bepalen van de totale tijd die hij of zij moet wachten. We kunnen bijgevolg de stochastiek $Z = X + Y$ definiëren. Hoe is deze stochastiek verdeeld ? En hoe is de stochastiek $Z = \sum_{i=1}^{n} X_i$ verdeeld waar elk stochastiek X_i de wachtijd weergeeft in eenzelfde tijdsinterval van 10 minuten ? Wat gebeurt er indien $n \to \infty$?

4.4 De normale of Gaussiaanse verdeling

Wanneer we een aantal onafhankelijke stochastische variabelen sommeren, dan gaat de waarschijnlijkheidsdichtheidsverdeling van de som steeds meer lijken op een zogenaamde normale verdeling (zie Hoofdstuk 5).

Dit is de basis van het belang van de normale verdeling, daar een waarneming vaak bestaat uit de som van een groot aantal toevalsgrootheden of stochastieken. In het geheel van alle theoretische verdelingen is de normale verdeling de belangrijkste. Ze wordt ook de Gaussiaanse verdeling genoemd naar Carl Friedrich Gauss (portret in Figuur 4.3), maar de vroegste vermelding vindt men in het werk van de engelsman Abraham de Moivre (1667-1754) in 1733. De meeste empirische gegevens volgen deze normale verdeling, vandaar de verwijzing naar 'normaal'.

Figuur 4.3: *Portret van Carl Friedrich Gauss (1777-1855).*

We hebben gezien dat de som Y van n onafhankelijke Bernoulli experimenten $\{X_i \mid i \in \{1, 2, ..., n\}\}$ met verdeling $B(1, p)$, binomiaal verdeeld is. De verwachtingswaarde van $Y \sim B(n, p)$ is bijgevolg gelijk aan np en de variantie gelijk aan $np(1-p)$. De genormaliseerde verdeling

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} \tag{4.26}$$

heeft bijgevolg een verwachtingswaarde gelijk aan 0 en een variantie gelijk aan 1. Indien $n \to \infty$ bekomen we een mooie symmetrische verdeling die in de limiet de volgende waarschijnlijkheidsdichtheidsverdeling aanneemt

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \tag{4.27}$$

welke we de standaard of genormeerde normale verdeling noemen en noteren met

$$Z \sim N(\mu = 0, \mathrm{Var} = 1) \tag{4.28}$$

waar de verwachtingswaarde en de variantie de curve volledig definiëren. Door over te gaan naar poolcoördinaten

$$\left\{ \begin{array}{lcl} x & = & r\,cos\phi \\ y & = & r\,sin\phi \end{array} \right. \tag{4.29}$$

kunnen we aantonen dat

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2}\,dx\right)^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2+y^2)}\,dxdy = \int_0^{2\pi}\int_0^{+\infty} e^{-\frac{1}{2}r^2}\,rdrd\phi = 2\pi \tag{4.30}$$

en dat bijgevolg de totale waarschijnlijkheid van de verdeling 1 is. Ook de verwachtingswaarde $E[Z]$ kunnen we verifiëren doordat de verdeling volledig symmetrisch is

$$E[Z] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} z\,e^{-\frac{1}{2}z^2} = \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{0} z\,e^{-\frac{1}{2}z^2}dz + \int_0^{+\infty} z\,e^{-\frac{1}{2}z^2}dz\right) = 0\ . \tag{4.31}$$

De variantie kunnen we uitrekenen met partiële integratie

$$\int_{-\infty}^{+\infty} x^2\,e^{-\frac{1}{2}x^2}\,dx = -\int_{-\infty}^{+\infty} x\,de^{-\frac{1}{2}x^2} = -xe^{-\frac{1}{2}x^2}\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2}dx = \sqrt{2\pi} \tag{4.32}$$

zodat

$$\mathrm{Var}[Z] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} z^2\,e^{-\frac{1}{2}z^2}\,dz = 1\ . \tag{4.33}$$

We kunnen de verdeling ook neerschrijven voor een willekeurige verwachtingswaarde μ en variantie σ^2, namelijk

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\,exp\left[-\frac{1}{2}\,\left(\frac{z-\mu}{\sigma}\right)^2\right] \tag{4.34}$$

die we noteren met

$$Z \sim N(\mu,\sigma^2)\ . \tag{4.35}$$

Een belangrijk gegeven voor de volgende hoofdstukken is dat de waarschijnlijkheid dat de normaal verdeelde veranderlijke X een waarde x aanneemt, die meer dan $\sigma_x = \sqrt{\mathrm{Var}[X]}$ afwijkt van zijn verwachtingswaarde μ_x, ongeveer gelijk is aan $1/3$ of

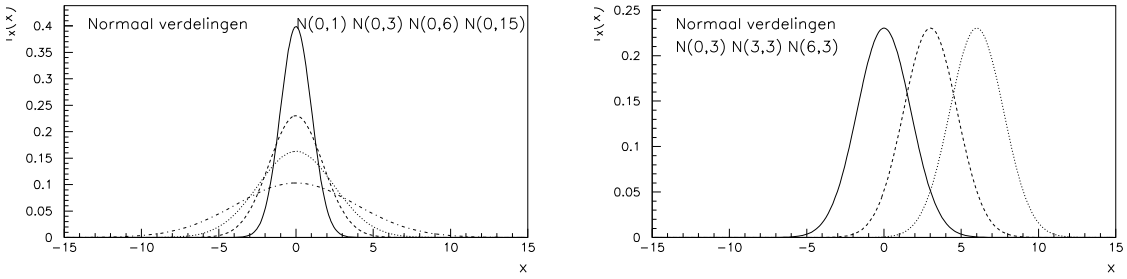
$$P(\,|x-\mu|\,<\sigma\,) = 0.68269 \simeq \frac{2}{3}\ . \tag{4.36}$$

Het buigpunt van de curve ligt ook op een afstand σ_x van de verwachtingswaarde μ_x, daar waar de afgeleide $f''_X(x)$ van teken verandert.

Een interessante stelling in verband met normaal verdeelde stochastieken, is dat hun som opnieuw normaal verdeeld is. Als $X \sim N(\mu_x,\sigma_x^2)$ en $Y \sim N(\mu_y,\sigma_y^2)$ onafhankelijk en normaal verdeeld zijn, dan is hun som $Z = X+Y$ opnieuw normaal verdeeld met verwachtingswaarde $\mu_z = \mu_x + \mu_y$ en variantie $\mathrm{Var}[Z] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ en bijgevolg $Z \sim N(\mu_z,\sigma_z^2)$.

Een unieke eigenschap voor de normale verdeling is dat indien $X_i \sim N(\mu, \sigma^2)$, het rekenkundig gemiddelde $\overline{x}$ en de steekproef variantie s^2 onafhankelijke grootheden zijn.

In Figuur 4.4 vindt men enkele voorbeelden van de waarschijnlijkheidsdichtheidsverdeling $f_X(x)$ van enkele stochastieken die een normale verdeling $N(\mu_x, \sigma_x^2)$ volgen.



Figuur 4.4: *Enkele voorbeelden van de continue normale of Gaussiaanse verdeling.*

De normale verdeling $f_Z(z)$ kan men uitbreiden naar meerdere dimensies $f_{\vec{Z}}(\vec{z})$. Om deze meer-dimensionale verdeling te beschrijven, gebruikt men matrix notatie. In de volgende studiejaren zullen we het nut en gemak inzien van dergelijke compacte notatie.

Als $\{X_i \sim N(0,1) \mid i \in \{1,2,...,n\}\}$ onafhankelijk zijn, dan is de n-dimensionale waarschijnlijkheidsdichtheidsverdeling $f_{\vec{X}}(\vec{x}) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot ... \cdot f_{X_n}(x_n)$ constant op de n-dimensionale sfeer met middelpunt $\vec{x} = 0$ en straal $R \in \Re$, omdat de verdeling enkel functie is van $R = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$. Ook dit is een eigenschap uniek aan de normale verdeling. Werk dit zelf uit!

4.5 De exponentiële verdeling

Als we terugkomen op het voorbeeld van de verkeersongelukken, kunnen we ook een stochastische variabele T definiëren die de tijd weergeeft tussen twee verkeersongelukken. Deze stochastiek T kan alle waarden aannemen tussen 0 en ∞, maar kan nooit negatief zijn. Om de waarschijnlijkheidsdichtheid te bepalen, kunnen we het aantal botsingen X beschouwen die gebeuren gedurende een tijdsinterval $[t, t+\tau]$. De duur van dit tijdsinterval is gelijk aan τ minuten en we stellen het gemiddeld aantal botsingen per minuut gelijk aan λ. Bijgevolg volgt X een Poisson verdeling $P_X(\lambda\tau)$ en is de waarschijnlijkheid dat er in het tijdsinterval τ in totaal k botsingen gebeuren gelijk aan

$$P_X(k) = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau} \quad . \tag{4.37}$$

In het bijzonder is de waarschijnlijkheid dat er geen botsingen gebeuren gelijk aan

$$P_X(0) = \frac{(\lambda\tau)^0}{0!} e^{-\lambda\tau} = e^{-\lambda\tau} \quad . \tag{4.38}$$

De waarschijnlijkheid $P(T > \tau)$ dat er geen botsingen plaatsvinden gedurende een periode τ is bijgevolg

$$P(T > \tau) = P_X(0) = e^{-\lambda\tau} \tag{4.39}$$

zodat

$$P(T \leq \tau) = 1 - P(T > \tau) = 1 - e^{-\lambda\tau} \qquad \text{voor } \tau > 0 \ . \tag{4.40}$$

De stochastiek T heeft dus een continue waarschijnlijkheidsdichtheidsverdeling $f_T(t)$ die 0 is voor alle negatieve waarden van t en

$$f_T(t) = \lambda\, e^{-\lambda t} \qquad \text{voor } t > 0 \tag{4.41}$$

en we noemen de verdeling een exponentiële verdeling met parameter λ.

Een belangrijke eigenschap van deze verdeling is dat ze lijdt aan geheugenverlies. Daar

$$P(T > s + t_0 \mid T > t_0) = \frac{P(T > s + t_0 \ \text{ en } \ T > t_0)}{P(T > t_0)} = \frac{e^{-\lambda(s+t_0)}}{e^{-\lambda t_0}} = e^{-\lambda s} = P(T > s) \tag{4.42}$$

en bijgevolg begint op elke plaats van de T-as eenzelfde verdeling (na hernormalisatie !). Indien de verdeling opnieuw moet beginnen na een zekere tijd t_0 zal die exact dezelfde verdeling volgen als deze die hij op een ander tijdstip zou beginnen. De waarschijnlijkheden die men berekent zijn onafhankelijk van het begintijdstip t_0 van de verdeling. Of nog anders geformuleerd, als we geen gebeurtenis waarnemen tot aan tijdstip t_1, dan is de waarschijnlijkheid om in een hieropvolgende periode tot tijdstip $t_2 > t_1$ ook geen gebeurtenissen waar te nemen onafhankelijk van tijdstip t_1. Men kan bewijzen dat de exponentiële verdeling de enige verdeling is die deze vergeetachtige eigenschap heeft.

De exponentiële verdeling $f_T(t)$ met parameter λ heeft volgende verwachtingswaarde

$$E[T] = \int_0^{+\infty} \lambda t e^{-\lambda t}\, dt = -te^{-\lambda t}\Big|_0^{+\infty} + \int_0^{+\infty} e^{-\lambda t}\, dt = \frac{1}{\lambda} \tag{4.43}$$

en via de relatie

$$E[T^2] = \int_0^{+\infty} \lambda t^2 e^{-\lambda t}\, dt = -t^2 e^{-\lambda t}\Big|_0^{+\infty} + \int_0^{+\infty} 2te^{-\lambda t}\, dt = \frac{2}{\lambda^2} \tag{4.44}$$

bekomen we de variantie

$$\text{Var}[T] = E[T^2] - E[T]^2 = \frac{1}{\lambda^2} \ . \tag{4.45}$$

Een toepassing van de exponentiële verdeling vinden we in het radioactief verval van stoffen. De levensduur T van elk atoom in een materiaal is exponentieel verdeeld met parameter λ. Als $N = N(t)$ de hoeveelheid radioactief materiaal op tijdstip t voorstelt, dan hebben we

$$N(t) = N(t = 0)e^{-\lambda t} \ . \tag{4.46}$$

De waarschijnlijkheid dat een atoom vervalt vóór het tijdstip t wordt gegeven door de cumulatieve $F_T(t)$ van de waarschijnlijkheidsdichtheidsverdeling $f_T(t)$

$$F_T(t) = 1 - e^{-\lambda t} \tag{4.47}$$

wat ons onmiddelijk de relatie 4.46 geeft. Uit 4.46 volgt dat

$$\frac{dN(t)}{dt} = -\lambda N(t) \ . \tag{4.48}$$

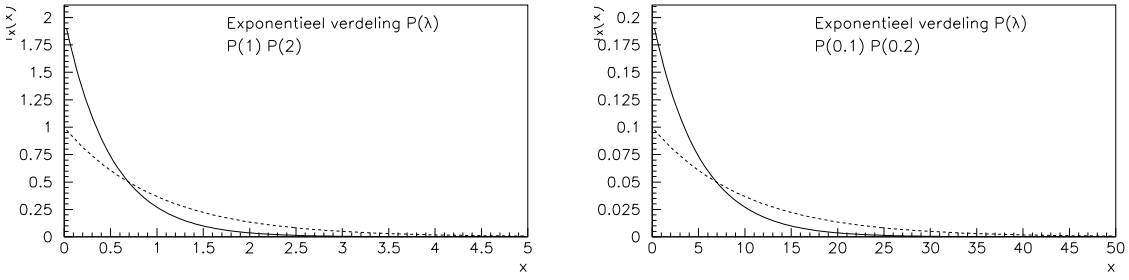
De mediaan voor deze verdeling is de tijd die nodig is om de hoeveelheid radioactief materiaal tot de helft te reduceren. Vandaar dat we dit de halfwaardetijd $t_{1/2}$ noemen en we vinden dat die gelijk is aan

$$t_{1/2} = \frac{ln2}{\lambda} \tag{4.49}$$

waarvoor geldt

$$F_T(t_{1/2}) = P(T \leq t_{1/2}) = \frac{N(t = t_{1/2})}{N(t = 0)} = \frac{1}{2} \ . \tag{4.50}$$

Let wel, deze verdeling gaat op voor een radioactieve atoomkern die op elk tijdstip t_0 aan eenzelfde verdeling begint voor zijn verval. Voor de levensduur van een mens is dit niet het geval. Een persoon van 70 jaar heeft gemiddeld een kortere levensverwachting dan iemand van 18 jaar. De verdeling van de levensduur is bijgevolg afhankelijk van het begintijdstip t_0. Daaraan zie je dat verzekeringsmaatschappijen sterk geïnteresseerd zijn in statistische begrippen !



Figuur 4.5: *Enkele voorbeelden van de continue exponenti¨ele verdeling.*

In Figuur 4.5 vinden we elke voorbeelden van de exponentiële verdeling $P(\lambda)$. Zoals we zien, kunnen we een verandering van λ_1 naar λ_2 interpreteren als een herschaling van de X-as (x) en Y-as ($f_X(x)$) op de grafiek. Bijgevolg zijn de waarschijnlijkheidsdichtheidsverdelingen van de stochastieken $X_1 \sim P(\lambda_1)$ en $X_2 \sim P(\lambda_2)$, namelijk $f_{X_1}(x)$ en $f_{X_2}(x)$, identiek op een herschaling van de X-as en een hernormalisatie na.

4.6 De chi-kwadraat verdeling

Als de stochastieken $\{X_i \mid i \in \{1, 2, ..., n\}\}$ onafhankelijk zijn en bovendien standaard normaal verdeeld zijn, met andere woorden $X_i \sim N(0, 1)$, dan heeft de som van de kwadraten

$$X = \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + ... + X_n^2 \tag{4.51}$$

een chi-kwadraat verdeling met n vrijheidsgraden en noteren we dit met

$$X \sim \chi_n^2 \ . \tag{4.52}$$

De waarschijnlijkheidsdichtheidsverdeling van stochastiek X wordt gegeven door

$$f_X(x) = \frac{1}{2^{\frac{n}{2}}\ \Gamma(\frac{n}{2})}\ x^{\frac{n}{2}-1}\ e^{-\frac{x}{2}} \tag{4.53}$$

voor $x \geq 0$ met de volgende definitie van de Γ-functie

$$\Gamma(t) = \int_0^{+\infty} x^{t-1}\ e^{-x}\ dx \tag{4.54}$$

en volgende eigenschappen

- $\Gamma(t+1) = t\Gamma(t)$;
- $\Gamma(n+1) = n!$ waar n een positief geheel getal is ;
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$;
- $\Gamma(\frac{n}{2}) = (\frac{n}{2} - 1)!$ als $n \geq 2$ en even is ;
- $\Gamma(\frac{n}{2}) = (\frac{n}{2} - 1)(\frac{n}{2} - 2)...(\frac{1}{2})\sqrt{\pi}$ als $n > 2$ en oneven is.

Met deze eigenschappen kunnen we de verwachtingswaarde bepalen

$$E[X] = \frac{1}{2^{\frac{n}{2}}\ \Gamma(\frac{n}{2})}\ \int_0^{+\infty} x\ e^{-\frac{x}{2}}\ x^{\frac{n}{2}-1}\ dx = \frac{2}{\Gamma(\frac{n}{2})} \int_0^{+\infty} e^{-\frac{x}{2}}\ \left(\frac{x}{2}\right)^{\frac{n}{2}}\ d\frac{x}{2} = \frac{2\Gamma(\frac{n}{2}-1)}{\Gamma(\frac{n}{2})} = n \tag{4.55}$$

en de variantie

$$\mathrm{Var}[X] = E[(X - E[X])^2] = E[X^2] + n^2 - 2nE[X] = E[X^2] - n^2 \ . \tag{4.56}$$

Wetende dat

$$E[X^2] = \frac{4\Gamma(\frac{n}{2}+2)}{\Gamma(\frac{n}{2})} = (n+2)n \tag{4.57}$$

vinden we dat

$$\mathrm{Var}[X] = 2n \ . \tag{4.58}$$

Als twee stochastische variabelen X en Y een χ^2-verdeling volgen, namelijk $X \sim \chi^2_{n_1}$ en $Y \sim \chi^2_{n_2}$, met respectievelijk n_1 en n_2 vrijheidsgraden, dan volgt de stochastische variabele $Z = X + Y$ eveneens een χ^2-verdeling met $n_1 + n_2$ vrijheidsgraden. We kunnen namelijk stellen dat

$$X = X_1^2 + X_2^2 + ... + X_{n_1}^2 \tag{4.59}$$

en

$$Y = Y_{n_1+1}^2 + Y_{n_1+2}^2 + ... + Y_{n_1+n_2}^2 \tag{4.60}$$

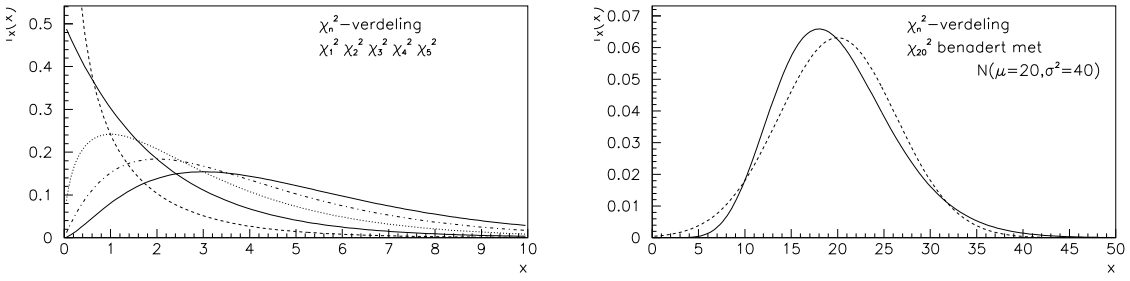
en bijgevolg

$$Z = X + Y = X_1^2 + X_2^2 + ... + X_{n_1}^2 + Y_{n_1+1}^2 + Y_{n_1+2}^2 + ... + Y_{n_1+n_2}^2 \sim \chi^2_{n_1+n_2} \ . \tag{4.61}$$

Ook de χ^2_n-verdeling wordt benaderd door een normale verdeling indien n groot wordt. De genormeerde stochastische variabele Z met

$$Z = \frac{\chi^2_n - n}{\sqrt{2n}} \tag{4.62}$$

convergeert naar de normale verdeling $N(0,1)$ indien $n \to \infty$. Dit is reeds een goede benadering indien $n \geq 30$.



Figuur 4.6: *Enkele voorbeelden van de continue χ^2_n-verdeling voor verschillende vrijheidsgraden n en een benadering met een normale verdeling $N(\mu, \sigma^2)$ aangeduid met een stippellijn.*

In Figuur 4.6 vinden we enkele voorbeelden van de χ^2_n-verdeling. We zien dat deze niet symmetrisch is en dat de verdeling goed benadert wordt door een normale verdeling indien het aantal vrijheidsgraden groot wordt.

De χ^2_n-verdeling heeft veel interessante eigenschappen. Eén ervan wordt duidelijk door volgende redenering. Stel dat n stochastische variabelen X_i onafhankelijk en normaal verdeeld

$X \sim N(0,1)$ zijn. We kunnen de veranderlijken voorstellen door een punt in de n-dimensionale steekproefruimte Ω. Omdat we met onafhankelijke stochastieken werken, is de waarschijnlijkheid om een punt in die ruimte aan te treffen evenredig met het product van de individuele normaal verdeelde waarschijnlijkheden en bijgevolg evenredig met $e^{-\frac{1}{2}\chi_n^2}$. De punten met gelijke waarschijnlijkheden liggen op een (n-1)-dimensionale hyperbol met straal χ_n. De waarschijnlijkheid voor een waarde van χ_n tussen χ_n en $\chi_n + d\chi_n$ is dus evenredig met $e^{-\frac{1}{2}\chi_n^2}$, vermenigvuldigd met het volume van de hyperbol-schil. Dit laatste is evenredig met $\chi_n^{n-1} d\chi_n$ zodat

$$P(\chi_n) \sim \chi_n^{n-1}\, e^{-\frac{1}{2}\chi_n^2}\, d\chi_n \tag{4.63}$$

We zoeken echter $P(\chi_n^2)$, bijgevolg

$$P(\chi_n^2) = P(\chi_n)\frac{d\chi_n}{d\chi_n^2} \sim \chi_n^{n-1}\, e^{-\frac{1}{2}\chi_n^2}\, \frac{1}{\chi_n}\, d\chi_n = \chi_n^{n-2}\, e^{-\frac{1}{2}\chi_n^2}\, d\chi_n \tag{4.64}$$

waar we de evenredigheidsconstante kunnen vinden door de totale waarschijnlijkheid te normaliseren op 1. Dit leidt tot het vroegere resultaat 4.53. Veronderstel nu dat er tussen de stochastische variabelen X_i een lineair verband bestaat

$$C_1 x_1 + C_2 x_2 + ... + C_n x_n = 0 \tag{4.65}$$

waarbij we de constanten C_i als reële getallen definiëren. De punten $\vec{x}$ liggen nu op een (n-1)-dimensionale sub-ruimte van de oorspronkelijke n-dimensionale steekproefruimte Ω. De waarschijnlijkheidsdichtheid is nog steeds evenredig met $e^{-\frac{1}{2}\chi_n^2}$. De vorige redenering is bijgevolg nog steeds geldig en men vindt dezelfde uitdrukking voor de waarschijnlijkheidsdichtheid, behalve dat de dimensie van de hyperbol met één is verminderd. Het aantal vrijheidsgraden is bijgevolg gereduceerd tot $n-1$.

We kunnen dit veralgemenen : Indien er r lineaire betrekkingen bestaan tussen de n stochastische variabelen $X_i \sim N(0,1)$, dan volgt de som van de kwadraten een χ_{n-r}^2-verdeling met $n-r$ vrijheidsgraden.

In Hoofdstuk 7 zullen we de χ_n^2-verdeling gebruiken om de verdeling van de variantie te bestuderen.

Eén van de variabelen die een χ_n^2-verdeling volgt en die veel gebruikt wordt tijdens wetenschappelijke experimenten is de zogenaamde Root Mean Square variabele. Zoals de naam suggereert is deze RMS waarde gedefinieerd als

$$V_{RMS,n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{4.66}$$

waar de stochastiek X bijvoorbeeld een ruissignaal beschrijft die meestal een normale verdeling volgt $X \sim N(0, V_{RMS,0})$. De waarde van $V_{RMS,0}$ is gedefinieerd als de verwachtingswaarde van X^2 of het tweede moment van X, bijgevolg is $E[X^2] = V_{RMS,0}^2$. Als de metingen van de electrische ruis van een versterker niet te snel na elkaar genomen zijn, kan men veronderstellen dat de metingen onafhankelijk zijn. Met die voorwaarde geldt dat de RMS stochastiek een χ_n^2-verdeling volgt met een schaalfactor

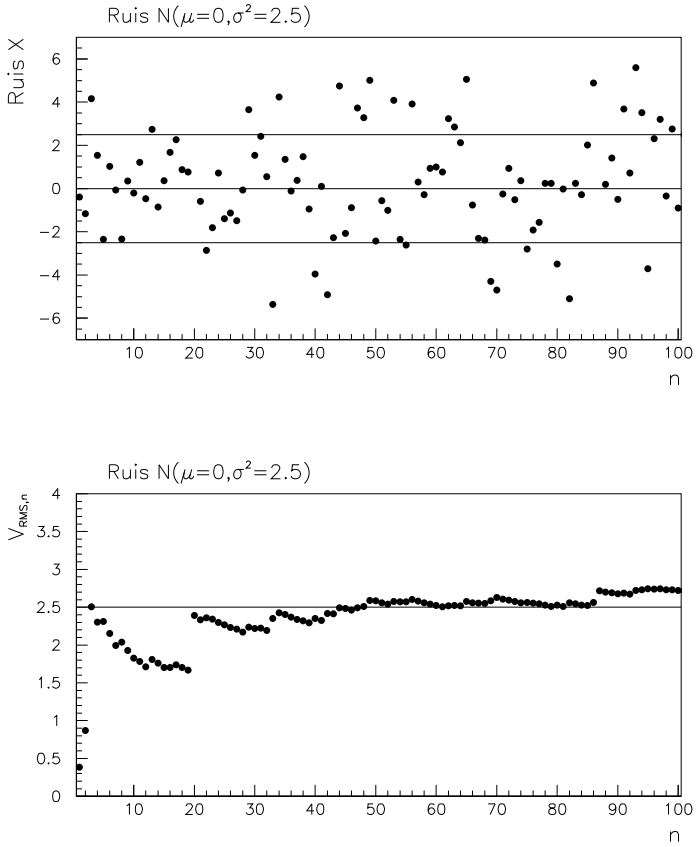
$$V_{RMS,n}^2 \sim \frac{V_{RMS,0}^2}{n}\,\chi_n^2 \quad . \tag{4.67}$$

Hiermee vinden we ook dat de verwachtingswaarde voor $V_{RMS,n}^2$ gelijk is aan

$$E[V_{RMS,n}^2] = \frac{V_{RMS,0}^2}{n}\,E[\chi_n^2] = V_{RMS,0}^2 \tag{4.68}$$

en dat indien het aantal metingen groot wordt ($n \to \infty$) de empirische waarde van $V_{RMS,n}^2$ convergeert naar de reële waarde, welke de grootte is van het ruissignaal $V_{RMS,0}^2$. Dit wordt verduidelijkt in Figuur 4.7 waar we 100 metingen doen van eenzelfde ruisverdeling en de empirische waarde van $V_{RMS,n}$ uitzetten. We zien dat de berekende waarde van $V_{RMS,n}$ convergeert naar $V_{RMS,0} = 2.5$ als we veel metingen maken. De zichtbare sprongen in de grafiek duiden op een uitschieter in de metingen die door het kwadraat een belangrijke invloed heeft. Het effect van die uitschieters wordt kleiner als het aantal metingen toeneemt.

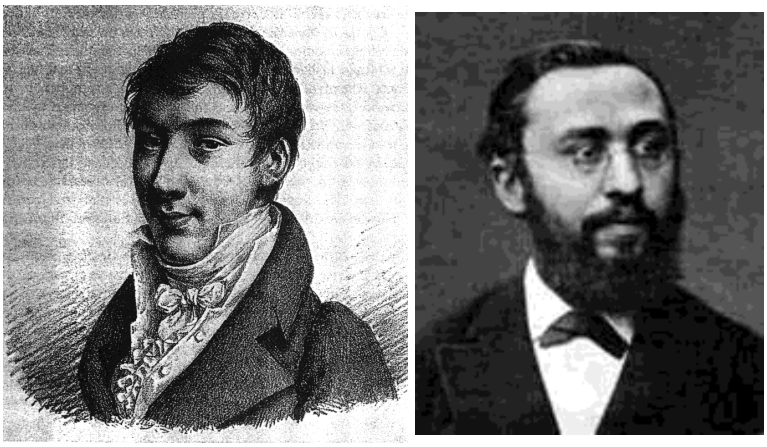
De Root Mean Square of RMS grootheid wordt bijvoorbeeld ook gebruikt om wisselstromen $I(t)$ te karakteriseren.



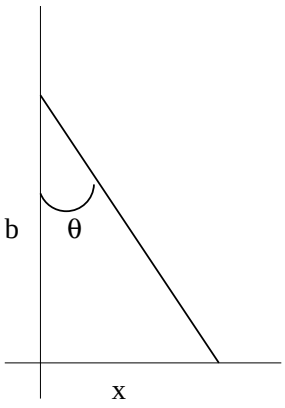
Figuur 4.7: *Een voorbeeld van 100 onafhankelijke metingen van ruis* $X \sim N(0, V_{RMS,0} = 2.5)$ *(links) en de empirische waarde van* $V_{RMS,n}$ *die we berekenen uit de eerste* n *metingen.*

4.7 De Cauchy verdeling

De Cauchy verdeling is een continue waarschijnlijkheidsdichtheidsverdeling die een resonantie beschrijft. De verdeling wordt genoemd naar Augustin-Louis Cauchy (1789-1857) en soms ook naar de Nederlander Hendrik Antoon Lorentz (1853-1928) als de Lorentz verdeling (zie Figuur 4.8 voor hun portretten). In de Natuurkunde kan deze verdeling ook de naam Breit-Wigner verdeling krijgen.



Figuur 4.8: *Portretten van Augustin-Louis Cauchy (links) en Hendrik Antoon Lorentz (rechts).*



Figuur 4.9: *Illustratie bij de afleiding van de Cauchy verdeling.*

De Cauchy verdeling beschrijft ook de verdeling van horizontale afstanden op de X-as die afgesneden worden door een lijn die een willekeurige hoek θ beschrijft met de Y-as, zie Figuur 4.9.

Stel de hoek θ gelijk aan de hoek die een rechte, met een vast rotatiepunt, maakt met de verticale Y-as. Dit kunnen we schrijven als

$$tan\theta = \frac{x}{b} \tag{4.69}$$

met de notaties zoals op de Figuur 4.9. Bijgevolg ook

$$\theta = tan^{-1}\left(\frac{x}{b}\right) \tag{4.70}$$

en afleiden naar de veranderlijken geeft

$$d\theta = -\frac{b\,dx}{b^2 + x^2} \quad . \tag{4.71}$$

De genormaliseerde verdeling van de hoek θ wordt dan

$$\frac{d\theta}{\pi} = -\frac{1}{\pi}\frac{b\,dx}{b^2+x^2} \tag{4.72}$$

daar de totale waarschijnlijkheid gelijk aan 1 is of

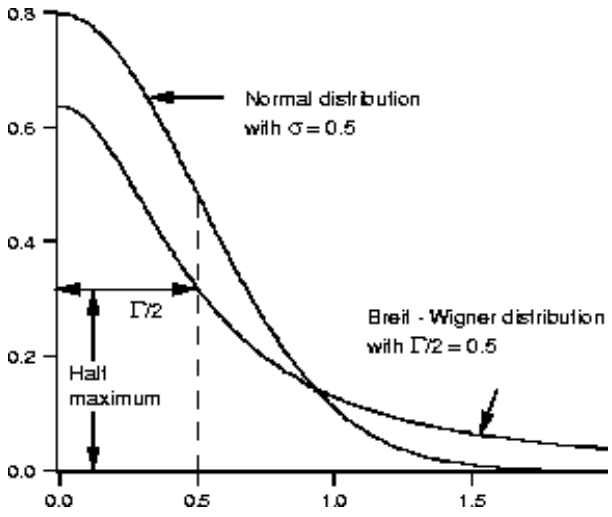
$$\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \frac{d\theta}{\pi} = 1 \ . \tag{4.73}$$

De algemene Cauchy verdeling kunnen we schrijven als

$$f_X(x) = \frac{1}{\pi}\frac{\frac{1}{2}\Gamma}{(x-m)^2 + \left(\frac{1}{2}\Gamma\right)^2} \tag{4.74}$$

waar Γ de volle breedte van de waarschijnlijkheidsdichtheidsverdeling weergeeft op halve hoogte [4] (of $\Gamma = 2b$ in bovenstaande afleiding) en m de mediaan is (of $m = 0$ in bovenstaande afleiding).

De momenten van deze verdeling kan men niet definiëren omdat de integralen divergeren. Daardoor neemt men de mediaan als maat voor de verwachtingswaarde en de breedte op halve hoogte als maat voor de spreiding. Men kan nagaan dat de som van Cauchy verdeelde onafhankelijke stochastieken terug een Cauchy verdeelde stochastiek is en dat indien X en Y normaal verdeelde stochastieken zijn met eenzelfde verwachtingswaarde, hun verhouding $Z = X/Y$ een Cauchy verdeling volgt met mediaan $m = 0$ en breedte op halve hoogte $\Gamma = 2\sigma_y/\sigma_x$.

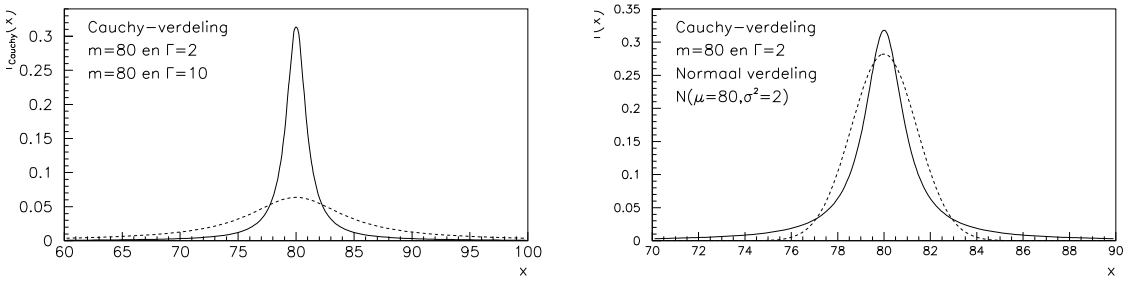


Figuur 4.10: *Vergelijking van een Cauchy en een normale verdeling.*

Een Cauchy verdeling wordt bijgevolg volledig gekarakteriseerd door een mediaan m en een breedte op halve hoogte Γ. In de studie van de elementaire deeltjes fysica worden deze grootheden geïnterpreteerd als een massa, m, en een breedte, Γ, van een resonant deeltje. Zo heeft het zogenaamde W boson een massa van ongeveer 80 GeV/c^2 en een breedte van ongeveer 2 GeV/c^2.

In Figuur 4.11 vinden we enkele voorbeelden van de Cauchy verdeling. We merken dat de normaal verdeling een benadering is van de Cauchy verdeling, maar dat de staarten van de normaal verdeling bijna geen waarschijnlijkheid bezitten, terwijl die staarten een niet verwaarloosbare waarschijnlijkheid hebben in een Cauchy verdeling. Dit wordt ook duidelijk door Figuur 4.10 te bestuderen. Van een normale verdeling kunnen we ook de

[4] In de literatuur vindt men soms ook de notatie FWHM voor deze grootheid, namelijk '*Full Width at Half Maximum*'.



Figuur 4.11: *Voorbeelden van enkele Cauchy verdelingen (links) en een benadering met een normaal verdeling (rechts).*

momenten berekenen daar de integralen convergeren, dit is, zoals reeds vermeld, niet het geval voor de Cauchy verdeling.

4.8 De Student-t verdeling

De verdeling geïntroduceerd door William Sealy Gosset (1876-1937) wordt over het algemeen de Student-t verdeling genoemd naar zijn pseudoniem. Een portret vinden we in Figuur 4.12.

Zijn werkgever, Guinness Breweries, verplichtte hem in 1908 om zijn werk te publiceren onder een pseudoniem en hij koos voor 'Student'.

Beschouw n onafhankelijke metingen x_i van een normaal verdeelde stochastiek $X \sim N(\mu, \sigma^2)$, dan kunnen we een t-waarde van de steekproef definiëren als

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{4.75}$$

waar μ de verwachtingswaarde is, $\overline{x}$ het rekenkundig gemiddelde van de steekproef en s de standaardafwijking van de steekproef, zoals gedefinieerd in vergelijking 3.2.

De stochastiek T, die de variabele t beschrijft, volgt een Student-t verdeling met $n - 1$ vrijheidsgraden, wat we noteren met

$$T \sim t_{n-1} \ . \tag{4.76}$$

Dit impliceert dat er een verschillende Student-t verdeling is voor elke verandering van n.


'Student' in 1908

Figuur 4.12: *Portret van William Sealy Gosset (1876-1937).*

Het gebeurt dat men wil nagaan of een empirisch rekenkundig gemiddelde $\overline{x}$ overeenkomt met een theoretische waarde μ. In dit geval gebruikt met de stochastiek T om deze hypothese te testen [5]. Let erop dat de variantie niet moet gekend zijn, maar dat we de steekproef variantie bepalen via de empirische gegevens.

De waarschijnlijkheidsdichtheidsverdeling voor de stochastiek $T \sim t_n$ is symmetrisch en benadert sterk de normale verdeling indien n groot wordt

$$t_n \simeq N(0,1) \qquad \text{indien } \; n \to \infty \; . \tag{4.77}$$

De uitdrukking van de waarschijnlijkheidsdichtheidsverdeling $f_T(t)$ is

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \, \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \tag{4.78}$$

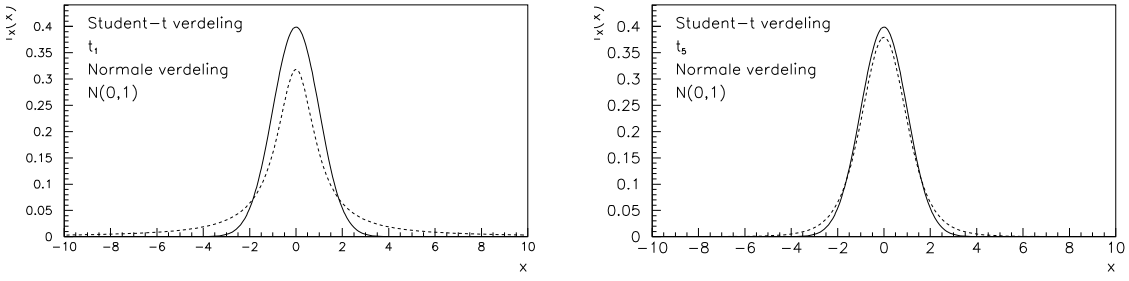
met een verwachtingswaarde die gelijk is aan 0 of

$$E[T] = 0 \tag{4.79}$$

en een variantie

$$\mathrm{Var}[T] = \frac{n}{n-2} \qquad \text{voor } \; n > 2 \; . \tag{4.80}$$

In Figuur 4.13 vinden we enkele voorbeelden van de Student-t verdeling, die reeds snel (voor $n > 5$) goed wordt benaderd door een standaard normale verdeling $N(0,1)$. Voor $n = 30$ is er bijna geen visueel onderscheid meer tussen beide verdelingen.



Figuur 4.13: *Voorbeelden van enkele Student-t verdelingen t_n en hun benadering met de standaard normale verdeling $N(0,1)$.*

Een alternatieve definitie voor de Student-t verdeling is de volgende. Beschouw een stochastiek X die een χ^2_n-verdeling volgt met n vrijheidsgraden ($X \sim \chi^2_n$) en een stochastiek Y die een standaard normaal verdeling volgt ($Y \sim N(0,1)$) en daarenboven onafhankelijk is van X, dan heeft de verhouding T

[5] De begrippen rond het testen van hypothesen worden in de cursus Statistiek van de volgende studiejaren Natuurkunde ge¨ı ntroduceerd.

$$T = \frac{Y}{\sqrt{\frac{X}{n}}} \tag{4.81}$$

een Student-t verdeling met n vrijheidsgraden ($T \sim t_n$). Via deze definitie is het eenvoudig na te gaan dat T een symmetrische verdeling heeft, daar beide stochastieken X en Y onafhankelijke zijn, en de stochastiek Y symmetrisch is. Men kan ook nagaan dat enkel de eerste $n-1$ momenten bestaan, daar de integralen voor de hogere momenten divergeren.

Hoofdstuk 5

Limietstellingen

"You can never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant.'

A.C. Doyle,
Sherlock Holmes in 'The Sign of Four'

In de praktijk maken we meestal verschillende metingen van eenzelfde stochastiek. Het is daarom belangrijk na te gaan hoe de waarschijnlijkheidsdichtheidsverdeling van een lineaire combinatie van dergelijke stochastieken eruit ziet. Met andere woorden, hoe kunnen we onze stochastiek beschrijven in de limiet van oneindig veel of althans héél veel metingen?

5.1 Algemeen

In de waarschijnlijkheidsrekening alsook in de statistiek bestaan er verschillende limietstellingen. Wij zullen hier enkel de belangrijkste beschouwen. Vele andere stellingen ontmoet je waarschijnlijk in de cursussen Differentiaal- en Integraalrekening, maar dan misschien uitgaande van een andere doelstelling.

Beschouw een stochastiek X met verwachtingswaarde μ_x en spreiding σ_x die beschreven wordt door een willekeurige verdeling. De ongelijkheid van Pavnutii Lvovich Tchebycheff (1821-1894) (portret in Figuur 5.1) zegt dat de waarschijnlijkheid dat X een waarde x zal aannemen die meer dan k maal de spreiding σ_x afwijkt van de verwachtingswaarde μ_x, kleiner is dan $1/k^2$ of

$$P(|x - \mu_x| > k\sigma_x) < \frac{1}{k^2} \ . \tag{5.1}$$

Voor een continue verdeling $f_X(x)$ wordt de variantie $\mathrm{Var}[X]$ gegeven door

$$\mathrm{Var}[X] = \sigma_x^2 = \int_{-\infty}^{+\infty} (x - \mu_x)^2 \ f_X(x) \ dx \tag{5.2}$$

of indien we de integraal opsplitsen

$$\sigma_x^2 = \int_{-\infty}^{\mu_x - k\sigma_x} (x-\mu_x)^2 f_X(x)dx + \int_{\mu_x - k\sigma_x}^{\mu_x + k\sigma_x} (x-\mu_x)^2 f_X(x)dx + \int_{\mu_x + k\sigma_x}^{+\infty} (x-\mu_x)^2 f_X(x)dx \tag{5.3}$$

zodat we de variantie kunnen afschatten met

$$\sigma_x^2 \; > \; \int_{-\infty}^{\mu_x - k\sigma_x} (x-\mu_x)^2 f_X(x)dx + \int_{\mu_x + k\sigma_x}^{+\infty} (x-\mu_x)^2 f_X(x)dx \quad . \tag{5.4}$$

In beide integralen kunnen we $(x-\mu_x)^2$ vervangen door haar kleinste waarde, namelijk $k^2\sigma_x^2$ zodat we het volgende bekomen

$$\sigma_x^2 \; > \; (k\sigma_x)^2 \left(\int_{-\infty}^{\mu_x - k\sigma_x} f_X(x)dx + \int_{\mu_x + k\sigma_x}^{+\infty} f_X(x)dx \right) \tag{5.5}$$

en bijgevolg

$$\sigma_x^2 \; > \; (k\sigma_x)^2 \, P(|x-\mu_x| > k\sigma_x) \tag{5.6}$$

waaruit de ongelijkheid van Tchebycheff 5.1 volgt.

En alternatieve manier om de ongelijkheid neer te schrijven is

$$P(|x-\mu_x| > \epsilon) < \frac{\sigma_x^2}{\epsilon^2} \tag{5.7}$$

waar men $k\sigma_x = \epsilon$ stelt in vorige ongelijkheid.

Een van de gevolgen van de ongelijkheid van Tchebycheff is de zogenaamde wet van de grote getallen. Beschouwen we een stochastiek Y met waarden y die het rekenkundig gemiddelde is van n onafhankelijke waarnemingen $\{x_i \mid i \in \{1, 2, ..., n\}\}$ van eenzelfde willekeurige stochastiek X met een eindige verwachtingswaarde μ_x en variantie $\mathrm{Var}[X] = \sigma_x^2$, bijgevolg

$$y = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.8}$$

en

$$E[Y] = \mu_x \qquad \mathrm{Var}[Y] = \sigma_y^2 = \frac{\sigma_x^2}{n} \tag{5.9}$$

Figuur 5.1: *Portret van Pavnutii Lvovich Tchebycheff (1821-1894).*

Dan convergeert de variantie van Y naar nul indien $n \to \infty$. Bijgevolg geldt via de ongelijkheid van Tchebycheff dat

$$P(|y - \mu_y| > \epsilon) < \frac{\sigma_y^2}{\epsilon^2} \tag{5.10}$$

en

$$P(|\overline{x} - \mu_x| > \epsilon) < \frac{1}{\epsilon^2}\frac{\sigma_x^2}{n} \stackrel{n\to\infty}{\longrightarrow} 0 \tag{5.11}$$

zodat de waarschijnlijkheid voor $\overline{x}$ convergeert naar μ_x.

Dit noemen we de wet van de grote getallen. De waarschijnlijkheid dat het rekenkundig gemiddelde $\overline{x}$ van n onafhankelijke waarnemingen van eenzelfde stochastiek X met verwachtingswaarde μ_x en variantie $\mathrm{Var}[X] = \sigma_x^2$, meer dan ϵ afwijkt van μ_x wordt willekeurig klein als $n \to \infty$.

Door het veelvuldig herhalen van een experiment kunnen we bijgevolg een fysische parameter van een verdeling willekeurig nauwkeurig bepalen.

5.2 De centrale limietstelling

Laat ons even terugkeren naar het bord van Galton. We kunnen het doorlopen van het bord van Galton beschouwen als een opeenvolging en bijgevolg een som van verschillende Bernoulli experimenten. De verdeling van de balletjes onderaan het bord of de som van de Bernoulli experimenten benadert een normale verdeling indien we meer rijen paaltjes toevoegen. Hoe meer rijen we toevoegen, hoe beter de benadering.

Deze observatie kunnen we veralgemenen in de zogenaamde centrale limietstelling, welke misschien de belangrijkste stelling is in de ganse cursus. Beschouw verschillende onafhankelijke stochastieken X_i die elk een willekeurige verdeling volgen. Elke stochastiek Y die men kan definiëren als een lineaire combinatie van deze stochastieken

$$Y = \sum_{i=1}^{n} c_i X_i \tag{5.12}$$

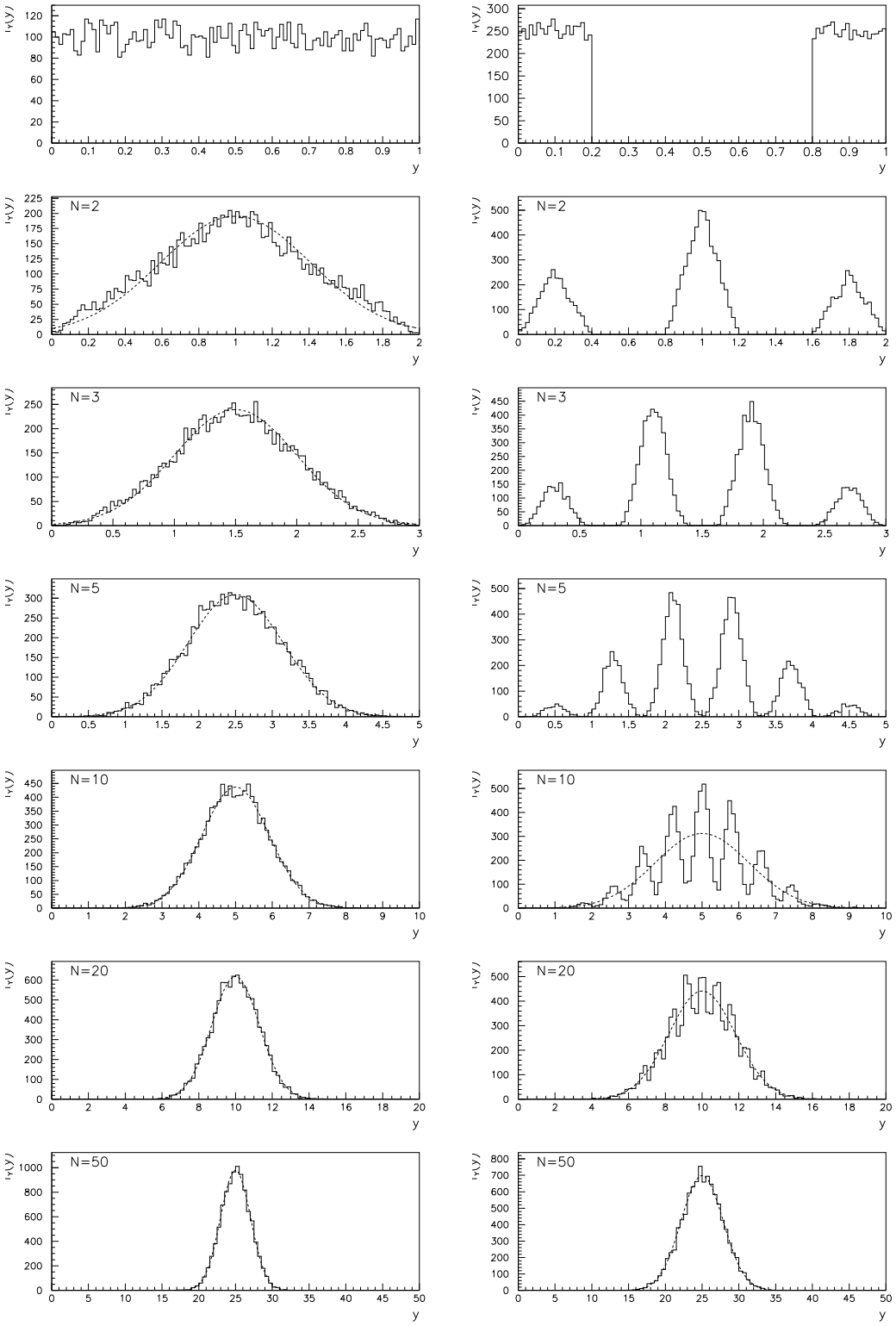
heeft ongeveer een normale verdeling. Om te vermijden dat één enkele stochastiek X_i de variantie van de som domineert, moeten we de voorwaarde invoeren dat de variantie

$$\mathrm{Var}[Y] = \sigma_y^2 = \sum_{i=1}^{n} c_i^2 \sigma_{x_i}^2 \tag{5.13}$$

veel groter is dan elk van de individuele componenten $c_i^2\sigma_{x_i}^2$. De benadering wordt beter, indien het aantal metingen n toeneemt

$$Y \;\sim\; N\left(\sum_{i=1}^{n} c_i E[X_i], \sum_{i=1}^{n} c_i^2 \sigma_{x_i}^2\right) \qquad \text{als } n \to \infty \;. \tag{5.14}$$

Door deze stelling kunnen we aannemen dat de theorie van de onzekerheden (meestal 'foutentheorie' genoemd) op een normale verdeling mag gebaseerd worden. Het concept van onzekerheden op het bepalen van een fysische grootheid is het onderwerp van Hoofdstuk 6. Het bewijs van de stelling kan men in de literatuur vinden, maar behoort niet tot de inhoud van deze cursus.



Figuur 5.2: *Illustratie van de centrale limietstelling voor twee verdelingen.*

In Figuur 5.2 vinden we twee stochastieken X_1 en X_2 waarmee we de centrale limietstelling kunnen illustreren. De eerste stochastiek X_1 beschrijft een uniforme verdeling in het interval $x_1 \in [0, 1]$, de tweede stochastiek X_2 beschrijft ook een uniforme verdeling maar de waarschijnlijkheid is gespreid over twee intervallen, namelijk $x_2 \in [0, 0.2]$ en $x_2 \in [0.8, 1]$. De histogrammen zijn gevuld met 10000 gesimuleerde gebeurtenissen. De bovenste histogrammen geven de empirische gegevens weer voor de stochastieken X_1 en X_2. In de daaropvolgende histogrammen worden de stochastieken $Y_{N,1}$ en $Y_{N,2}$ voorgesteld

$$Y_{N,1} = \sum_{i=1}^{N} X_{i,1} \tag{5.15}$$

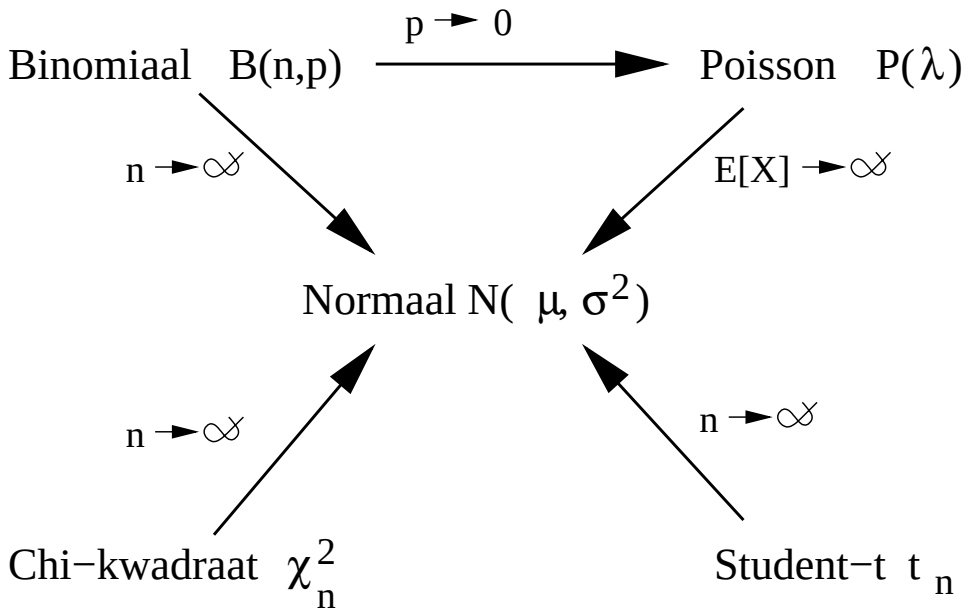
$$Y_{N,2} = \sum_{i=1}^{N} X_{i,2} \tag{5.16}$$

die een lineaire combinatie (hier de som of $c_i = 1$ voor alle i) zijn van verschillende gelijke stochastieken $X_{i,1} \sim X_1$ of van verschillende gelijke stochastieken $X_{i,2} \sim X_2$. De stippellijn toont de benadering met een normale verdeling volgens de centrale limietstelling 5.14.

We zien dat voor een combinatie van N empirische metingen van een stochastiek X_1 die een uniforme verdeling volgt in één interval, de benadering met de normale verdeling reeds goed is voor $N = 3$. Voor de bizarre stochastiek X_2 is dit slechts het geval voor $N > 20$.

In Hoofdstuk 6 zullen we de centrale limietstelling toepassen op het rekenkundig gemiddelde van een verzameling empirische gegevens.

We hebben één uitzondering gezien op de centrale limietstelling, namelijk de Cauchy verdeling. Het rekenkundig gemiddelde van n waarnemingen van een stochastiek, die beschreven wordt met een Cauchy verdeling, is ook verdeeld volgens een Cauchy verdeling met dezelfde parameters als de oorspronkelijke stochastiek.



Figuur 5.3: *Illustratie van de onderlinge verbanden tussen de standaard verdelingen.*

De meeste verdelingen, besproken in Hoofdstuk 4, hebben een onderling verband. Dit wordt geïllustreerd in Figuur 5.3. Verschillende verdelingen hebben een asymptotisch verband, en bijna allen hebben een asymptotisch verband met de normale verdeling.

De toepassingen van de centrale limietstelling zullen verder duidelijk worden tijdens het uitvoeren van een experiment. Om het brandstofverbruik te bepalen van een voertuig, moet men rekening houden met verschillende factoren. Men kan het traject van het voertuig indelen in verschillende intervallen. Het eerste interval (bijvoorbeeld in een dorpskern) heeft een lengte die beschreven wordt door stochastiek X_1, terwijl het tweede stukje (bijvoorbeeld op de snelweg) een lengte heeft die beschreven wordt door stochastiek X_2, enzovoort. Uiteindelijk heeft men n stukjes die elk hun eigen stochastiek hebben X_i. De totale lengte is een lineaire combinatie, namelijk de som, van alle stochastieken X_i en zal bijgevolg een normale verdeling volgen, indien we vele verschillende stukjes hebben of $n \to \infty$.

We kunnen ook terug het voorbeeld van de wachtende professor aanhalen. Als hij op elke student een willekeurige tijd tussen 0 en 1 minuut moet wachten, zal de total wachtijd voor N opeenvolgende studenten een waarschijnlijkheidsdichtheidsverdeling volgen die eruit ziet als in Figuur 5.2 (links). Hiervoor moeten we de histogrammen nog normaliseren.

Hoofdstuk 6

Parameter schatter en onzekerheden

"Everybody believes in the law of errors,
the experimenters because they think it can be proved
by mathematics, the mathematicians because they
believe it has been established by observation.''

G. Lippman (1845-1921),
in een brief aan Poincaré (1917)

Eén van de essentiële betrachtingen van de wetenschap is het opstellen van theorieën of modellen die de realiteit beschrijven. Deze nieuwe ideeën worden hieropvolgend getest door het verzamelen van experimentele gegevens over eigenschappen of karakteristieke grootheden van deze theorieën. Zo kan men van de relativiteitstheorie van Albert Einstein zeggen dat die met een zeer grote nauwkeurigheid geverifieerd is, terwijl we over de zogenaamde snaartheorie ('String Theory') slechts weinig experimentele bevestiging hebben. We zeggen dat we aan de eerste theorie meer geloof hechten dan aan de tweede. Om ons geloof of vertrouwen in een theorie te verhogen, moeten we experimentele gegevens verzamelen. Een theorie die een waarde van θ_0 voorspelt voor een fysische grootheid of een parameter, kan men verifiëren via een zogenaamde schatter $\widehat{\theta}$ die experimenteel meetbaar of observeerbaar is. In dit hoofdstuk gaan we bijgevolg uit van een verzameling van empirische gegevens.

6.1 Definitie en eigenschappen van een schatter

Voor een oneindig aantal metingen van een grootheid X kan men de exacte waarschijnlijkheidsdichtheidsverdeling $f_X(x)$ opstellen. In vorig hoofdstuk hebben we gezien hoe we van dergelijke verdelingen de momenten, zoals verwachtingswaarde μ_x en variantie σ_x^2, bepalen. Indien men slechts n empirische gegevens heeft, kunnen we deze beperkte steekproef van de volledige populatie karakteriseren aan de hand van grootheden zoals het rekenkundig gemiddelde $\overline{x}$, de variantie s_x, alsook alle andere momenten en centrale momenten [1]. Ook willen

[1] Let op, deze empirische grootheden noteren we met Latijnse symbolen en de theoretische grootheden, bekomen uit het theoretisch voorschrift van de verdeling, met Griekse symbolen.

we met behulp van deze beperkte steekproef de eigenschappen afleiden voor de volledige populatie, of nagaan wat de Latijnse symbolen kunnen zeggen over de Griekse symbolen. Hiervoor moeten we schatters $\widehat{\theta}$ opstellen die deze theoretische eigenschappen van de onderliggende verdeling benaderen. Schatters van de theoretische parameter θ zullen we aanduiden met $\widehat{\theta}$, de exacte constante waarde van de parameter met θ_0.

Merk op dat de theoretische momenten van de onderliggende verdeling niet onderhevig zijn aan statistische fluctuaties, maar de karakteristieken van de steekproef wel. Men heeft namelijk een willekeurige eindige verzameling van n metingen *at random* gekozen uit de totale oneindige verzameling van mogelijke metingen.

Zoals het woord schatten aanduidt, gaat het hier om een methode die niet precies is en bijgevolg niet tot exacte resultaten leidt. Daar we nooit oneindig veel metingen kunnen doen, moeten we altijd via schatters iets te weten komen over de onderliggende theoretische verdeling en zijn eigenschappen. We kunnen met andere woorden nooit exact meten, er blijft altijd een onzekerheid over omdat we geen oneindig aantal metingen kunnen doen. Wel hebben we in Hoofdstuk 5 aangetoond dat we de waarde van elke fysische grootheid willekeurig dicht kunnen benaderen, zodat de onzekerheid willekeurig klein wordt (zie ongelijkheid 5.11).

Stel dat de theoretische verdeling van stochastiek X afhankelijk is van de fysische parameters $\{\theta_1, \theta_2, ..., \theta_k\}$ die de constante waarden $\{\theta_{1,0}, \theta_{2,0}, ..., \theta_{k,0}\}$ aannemen. We kunnen de verdeling van stochastiek X noteren met $f_X(x \mid \theta_{1,0}, \theta_{2,0}, ..., \theta_{k,0})$. Dit is de waarschijnlijkheid voor x indien de parameters een zekere waarde aannemen. We willen nu een schatter $\widehat{\theta}_j$ opstellen voor parameter θ_j die zal afhangen van de n empirische gegevens $\{x_1, x_2, ..., x_n\}$, bekomen uit de steekproef. De waarde van de schatter $\widehat{\theta}_j$ zal bijgevolg een functie zijn van de empirische gegevens of

$$\widehat{\theta}_j = \widehat{\theta}_j(x_1, x_2, ..., x_n) \tag{6.1}$$

en indien men een andere steekproef neemt, zal men bijgevolg een andere waarde voor de schatter bekomen. Daar de schatter een functie is van stochastische variabelen, is hij zelf ook een stochastische variabele met een zekere waarschijnlijkheidsdichtheidsverdeling. De stochastiek $\widehat{\theta}_j$ die waarden θ_j aanneemt heeft een waarschijnlijkheidsdichtheidsverdeling die theoretisch berekend kan worden uit de theoretische verdeling van de stochastiek X. Laat ons voor de eenvoud $k = 1$ stellen, dan kunnen we spreken over een schatter $\widehat{\theta}$ die waarden θ aanneemt en gebruikt wordt om een parameter met constante waarde θ_0 te schatten.

Een schatter heeft dus een numerieke waarde en kan bijgevolg niet afhankelijk zijn van onbekende parameters, zoals de theoretische verwachtingswaarde μ_x of variantie σ_x^2 van de onderliggende verdeling van stochastiek X.

Een schatter is gedefinieerd als een functie van de n empirische gegevens x_i die bijgevolg een waarde heeft die we een schatting noemen. Er zijn vier belangrijke eigenschappen die de kwaliteit van een schatter definiëren:

- **Consistentie**: Een schatter $\widehat{\theta}$ wordt consistent genoemd, indien zijn waarde convergeert naar de waarde van de theoretische parameter θ_0 als het aantal metingen n toeneemt ($n \to \infty$).

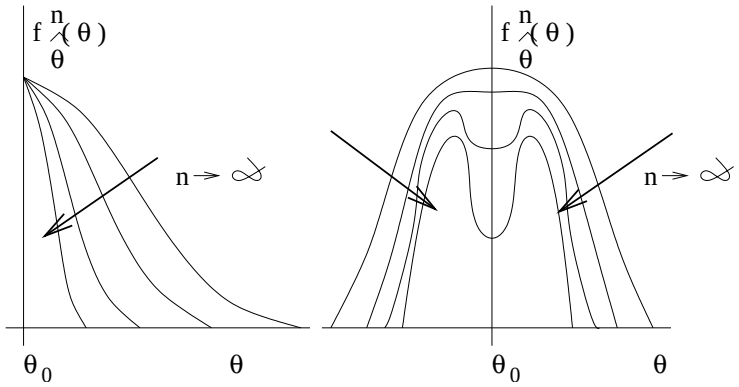
- **Zuiverheid**: Een schatter $\widehat{\theta}$ wordt zuiver genoemd indien de verwachtingswaarde van de schatter gelijk is aan de theoretische waarde van de parameter θ_0, bijgevolg $E[\widehat{\theta}] =$

θ_0. Indien deze voorwaarde niet geldt, zeggen we dat de schatter een verschuiving of *bias* heeft. Die bias noteren we als $b(\widehat{\theta})$ en is eigen aan de schatter $\widehat{\theta}$ en kan eventueel afhankelijk zijn van het aantal metingen n. Een schatter is bijgevolg zuiver indien $b_n(\widehat{\theta}) = 0$ voor alle waarden van n en θ_0.

- **Efficiëntie**: Een schatter $\widehat{\theta_1}$ is efficiënter dan een andere schatter $\widehat{\theta_2}$ indien zijn waarschijnlijkheidsdichtheidsverdeling $f_{\widehat{\theta_1}}(\theta_1)$ een kleinere variantie heeft, bijgevolg indien $\sigma^2_{\widehat{\theta_1}} < \sigma^2_{\widehat{\theta_2}}$.

- **Robuustheid**: Een schatter $\widehat{\theta}$ wordt robuust genoemd indien zijn numerieke waarden weinig afhankelijk zijn van uitschieters tussen de metingen x_i. Dit komt er ook op neer dat de schatter $\widehat{\theta}$ onafhankelijk moet zijn van de verdeling $f_X(x \mid \theta_0)$ of afwijkingen van deze verdeling.

Men zou denken dat er een eenduidig verband bestaat tussen de consistentie en de zuiverheid van een schatter. Maar we kunnen via tegenvoorbeelden aantonen dat indien de schatter voldoet aan de ene eigenschap, dit niet noodzakelijk impliceert dat ook aan de andere eigenschap voldaan wordt.



Figuur 6.1: *Illustratie bij de begrippen consistentie en zuiverheid.*

Denk maar aan een waarschijnlijkheidsdichtheidsverdeling voor de schatter $\widehat{\theta}$ die niet symmetrisch is. Asymptotisch ($n \to \infty$) kan een schatter zuiver zijn, maar niet noodzakelijk voor alle waarden van n. Voorbeeld in Figuur 6.1 (links). Deze waarschijnlijkheidsdichtheidsverdeling $f^n_{\widehat{\theta}}(\theta \mid \theta_0)$ van de schatter $\widehat{\theta}$ die waarden θ aanneemt, berekend uit n empirische metingen, is niet symmetrisch maar convergeert naar een δ-piek bij θ_0 indien n groot wordt ($n \to \infty$). Deze schatter is bijgevolg consistent, daar zijn verwachtingswaarde $E[\widehat{\theta}]$ convergeert naar θ_0 indien $n \to \infty$. De schatter is echter niet zuiver omdat de verwachtingswaarde steeds een bias vertoont, $b_n(\widehat{\theta}) \neq 0$ voor elke waarde van n behalve voor $n = \infty$.

Een waarschijnlijkheidsdichtheidsverdeling $f^n_{\widehat{\theta}}(\theta \mid \theta_0)$ van de schatter $\widehat{\theta}$ die waarden θ aanneemt, berekend uit n empirische metingen kan ook geen bias vertonen, maar toch niet consistent zijn. Dit wordt duidelijk in Figuur 6.1 (rechts), waar een verdeling symmetrisch convergeert naar twee δ-pieken bij verschillende waarden of $\theta_1 \neq \theta_2$. De verschuiving of

bias is bijgevolg steeds nul en de schatter is dus consistent, maar zijn waarde convergeert niet naar θ_0. Een reëel voorbeeld van een inconsistente schatter is het rekenkundig gemiddelde $\overline{x}$ voor een Cauchy verdeling. We hebben gezien dat de stochastiek $\overline{X}$ eenzelfde verdeling met dezelfde parameters volgt als de stochastiek X. Bijgevolg convergeert $\overline{X}$ eigenlijk naar niets.

Let ook op dat, indien $\widehat{\theta}$ een zuivere schatter is voor de parameter θ, dit niet impliceert dat $\widehat{\theta^2}$ een zuivere schatter is voor de parameter θ^2.

We beperken ons hier tot schatters voor de verwachtingswaarde $E[X] = \mu_x$ en de variantie $\mathrm{Var}[X] = \sigma_x^2$ van de theoretische onderliggende verdeling $f_X(x)$. De eerste grootheid is een kental voor de locatie van de verdeling, de tweede grootheid is een kental voor de spreiding van de verdeling.

6.2 Voorbeeld : het rekenkundig gemiddelde $\overline{x}$

Beschouw een verzameling van n empirische gegevens over de stochastiek X, die waarden x aanneemt en een waarschijnlijkheidsdichtheidsverdeling $f_X(x \mid \mu_0)$ heeft met een theoretische verwachtingswaarde μ, die een constante waarde μ_0 heeft. Omdat we de parameters van de theoretische verdeling niet kennen, moeten we die schatten uit de empirische gegevens. Om de locatie weer te geven van de empirische gegevens, bepalen we het rekenkundig gemiddelde

$$y = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \tag{6.2}$$

We kunnen bijgevolg $\overline{x}$ beschouwen als een lineaire combinatie van onafhankelijke stochastieken X_i. Indien n groot wordt, leren we uit de centrale limietstelling dat de stochastiek Y, die het rekenkundig gemiddelde beschrijft, een normale verdeling volgt. We kunnen het rekenkundig gemiddelde $\widehat{\mu} = \overline{x}$ beschouwen als een schatter voor de theoretische verwachtingswaarde μ_0, die een verdeling $f_{\widehat{\mu}}(\mu)$ volgt die de stochastiek $\widehat{\mu}$ beschrijft en waarden μ aanneemt.

We zien direct dat deze schatter zuiver is, namelijk

$$E[\widehat{\mu}] = E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n E[X] = E[X] = \mu_0 \tag{6.3}$$

en er bijgevolg geen verschuiving of bias, $b_n(\widehat{\mu}) = 0$ voor elke waarde $n \geq 1$..

Het rekenkundig gemiddelde is ook een efficiënte schatter, daar

$$\mathrm{Var}[\widehat{\mu}] = \mathrm{Var}[\overline{X}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{1}{n^2} n \mathrm{Var}[X] = \frac{\sigma_x^2}{n} \ . \tag{6.4}$$

De onzekerheid of spreiding op de variabele $\overline{X}$ is bijgevolg $1/\sqrt{n}$ keer kleiner dan de onzekerheid op de variabele X. De onzekerheid benadert nul indien n groot is. Hoe meer gegevens we nemen ($n \to \infty$), hoe kleiner de onzekerheid ($\mathrm{Var}[\widehat{\mu}] \to 0$). Dit is een nuttige uitdrukking indien we willen inschatten hoeveel keer we een meting moeten herhalen om een onzekerheid $\mathrm{Var}[\widehat{\mu}]$ te verkrijgen die kleiner is dan een zekere waarde.

In de meeste praktische gevallen kennen we de parameter σ_x^2 niet. Bijgevolg gebruikt men meestal de geschatte waarde s_x^2 (zie volgende sectie). Men bekomt dan de uitdrukking

$$\mathrm{Var}[\widehat{\mu}] = \frac{s_x^2}{n} \quad . \tag{6.5}$$

We kunnen bijgevolg stellen dat het rekenkundig gemiddelde $\overline{x}$ van n waarnemingen of metingen een goede schatter is voor de locatie van de verdeling $f_X(x)$.

Indien de waarden x_i een verschillende variantie $\sigma_{x_i}^2$ hebben, kunnen we een alternatieve schatter $\widehat{\mu_g}$ opstellen die een kleinere variantie heeft dan het rekenkundig gemiddelde. Dit door in de som een verschillend gewicht te geven aan iedere individuele meting, namelijk

$$\widehat{\mu_g} = \frac{1}{w}\sum_{i=1}^{n} w_i x_i \tag{6.6}$$

met

$$w_i = \frac{1}{\sigma_{x_i}^2} \qquad \text{en} \qquad w = \sum_{i=1}^{n} w_i \quad . \tag{6.7}$$

We noemen deze schatter, het gewogen gemiddelde. De variantie $\mathrm{Var}[\widehat{\mu_g}]$ is gelijk aan 1/w.

6.3 Voorbeeld : de steekproef variantie s_x^2

Beschouw opnieuw een verzameling van n empirische gegevens over de stochastiek X die waarden x aanneemt en een waarschijnlijkheidsdichtheidsverdeling $f_X(x \mid \sigma_x)$ heeft met een theoretische verwachtingswaarde μ die een constante maar ongekende waarde μ_0 heeft. We willen nu de variantie bepalen van stochastiek X. Hiervoor kunnen we volgende schatter nemen

$$y = \widehat{\sigma_x^2} = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu_x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \tag{6.8}$$

die we rechtstreeks kunnen berekenen via de n empirische gegevens.

We kunnen aantonen dat deze schatter zuiver is

$$E[\widehat{\sigma_x^2}] = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\right] \tag{6.9}$$

of

$$E[\widehat{\sigma_x^2}] = \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n} X_i^2\right] - \frac{1}{n}E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right]\right) \quad . \tag{6.10}$$

Indien de n metingen onafhankelijk zijn, geldt

$$E\left[\sum_{i=1}^{n} X_i^2\right] = nE[X^2] \tag{6.11}$$

en rekening houdend met de uitdrukkingen voor de theoretische verdelingen

$$\sigma_x^2 = E[X^2] - \mu_x^2 \tag{6.12}$$

en

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(E\left[\sum_{i=1}^{n} X_i\right]\right)^2 \tag{6.13}$$

bekomen we

$$E[\widehat{\sigma_x^2}] = \frac{1}{n-1}\left[n(\sigma_x^2 + \mu_x^2) - \frac{1}{n}\left(\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] + \left(E\left[\sum_{i=1}^{n} X_i\right]\right)^2\right)\right] \quad . \tag{6.14}$$

Via de eigenschappen van de verwachtingswaarde en de variantie bekomen we volgende vergelijkingen

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = n\mathrm{Var}[X] = n\sigma_x^2 \tag{6.15}$$

en

$$E\left[\sum_{i=1}^{n} X_i\right] = nE[X] = n\mu_x \tag{6.16}$$

die we kunnen gebruiken in vergelijking 6.14 om te bekomen dat

$$E[\widehat{\sigma_x^2}] = \frac{1}{n-1}\left[n\sigma_x^2 + n\mu_x^2 - \frac{1}{n}\left(n\sigma_x^2 + (n\mu_x)^2\right)\right] \tag{6.17}$$

en bijgevolg de voorwaarde voor zuiverheid

$$E[\widehat{\sigma_x^2}] = \frac{1}{n-1}(n-1)\sigma_x^2 = \sigma_x^2 \quad . \tag{6.18}$$

Om aan te tonen dat de schatter $\widehat{\sigma_x^2}$ zuiver is, hebben we nergens een voorwaarde geïntroduceerd voor de verdeling van de stochastiek X. Let erop dat, indien we als schatter voor σ_x^2 de volgende definitie nemen,

$$\widehat{\sigma_x^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu_x})^2 \tag{6.19}$$

niet aan deze voorwaarde zou voldaan zijn. Dit is één van de redenen waarom we in uitdrukking 3.2 delen door $(n-1)$ en niet door n. De correctie van n naar $n-1$ werd geïntroduceerd door Friedrich Wilhelm Bessel (1784-1846), maar kan ook toegeschreven worden aan Gauss die reeds in 1823 een gelijkaardige factor gebruikte.

We kunnen ook de variantie van de schatter $\widehat{\sigma_x^2}$ bepalen, namelijk

$$\mathrm{Var}[\widehat{\sigma_x^2}] = \mathrm{Var}\left[\frac{1}{n-1}\sigma_x^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\sigma_x^2}\right] = \left(\frac{\sigma_x^2}{n-1}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} z_i^2\right] \tag{6.20}$$

waar we Z gedefinieerd hebben als

$$Z = \frac{X - \overline{x}}{\sigma_x} \ . \tag{6.21}$$

We weten uit Hoofdstuk 4 dat $\sum_{i=1}^{n} z_i^2$ een χ^2_{n-1} verdeling volgt. Hiervan hebben we de variantie bepaald, namelijk

$$\mathrm{Var}\left[\sum_{i=1}^{n} z_i^2\right] = \mathrm{Var}\left[\chi^2_{n-1}\right] = 2(n-1) \tag{6.22}$$

en bijgevolg bekomen we voor de variantie van de schatter $\widehat{\sigma_x^2}$

$$\mathrm{Var}[\widehat{\sigma_x^2}] = \frac{2}{n-1}\sigma_x^4 \ . \tag{6.23}$$

We merken op dat deze uitdrukking de parameter σ_x bevat die in de meeste gevallen a priori niet gekend is. Het is bijgevolg gebruikelijk om de geschatte variantie $\widehat{\sigma_x^2} = s_x^2$ te gebruiken, zodat

$$\mathrm{Var}[\widehat{\sigma_x^2}] = \frac{2}{n-1}s_x^4 \ . \tag{6.24}$$

We kunnen bijgevolg stellen dat de steekproefvariantie s_x^2 een efficiënte schatter is voor de spreiding van de verdeling $f_X(x)$.

Men kan ook de onzekerheid op de onzekerheid van $\widehat{\mu_x}$ bepalen. De onzekerheid op $\widehat{\mu_x}$ is namelijk de geschatte standaardafwijking $\widehat{\sigma_{\mu_x}}$ die ook een stochatische verdeling volgt en die verdeling heeft bijgevolg ook een standaardafwijking, namelijk $\sigma(\widehat{\sigma_{\mu_x}})$. Dit kunnen we noteren met

$$\sigma(\sigma_{\mu_x}) = \sqrt{\mathrm{Var}[\sigma_{\mu_x}]} = \sqrt{\mathrm{Var}\left[\sqrt{\frac{s_x^2}{n}}\right]} = \sqrt{\frac{1}{n}\mathrm{Var}[s_x]} \tag{6.25}$$

waar we $\mathrm{Var}[s_x]$ kunnen bepalen uit

$$\mathrm{Var}[\widehat{\sigma_x^2}] = \mathrm{Var}[s_x^2] = \left(\frac{ds_x^2}{ds_x}\right)^2 \mathrm{Var}[s_x] = (2s_x)^2\mathrm{Var}[s_x] \tag{6.26}$$

of

$$\mathrm{Var}[s_x] = \frac{1}{2(n-1)}s_x^2 \tag{6.27}$$

en bijgevolg wordt vergelijking 6.25

$$\sigma(\sigma_{\mu_x}) = \sqrt{\frac{1}{2n(n-1)}} s_x = \frac{\sigma_{\mu_x}}{\sqrt{2n(n-1)}} \quad . \tag{6.28}$$

We vinden dat, zelfs als n relatief klein is, de onzekerheid op de onzekerheid op de verwachtingswaarde verwaarloosbaar is ten opzichte van de onzekerheid op de verwachtingswaarde. Bijgevolg is dit juist een academisch begrip en heeft dit geen impact op de resultaten van een experiment.

6.4 Interpretatie van de onzekerheid op de waarde van de schatter

Er bestaan verschillende methoden om een parameter te schatten. In Hoofdstuk 7 zullen we er één van nabij bekijken. Elk van deze methoden resulteert in een waarde θ voor de schatter $\widehat{\theta}$. Daar de waarde voor de schatter $\widehat{\theta}$ van een kengetal van de theoretische verdeling $f_X(x \mid \theta_0)$ niet gelijk is aan de theoretische waarde θ_0 van dit kengetal maar hiervan afwijkt, moeten we definiëren wat de waarschijnlijkheid is om de theoretische waarde θ_0 in de nabijheid van de geschatte waarde θ te vinden. Dit doen we aan de hand van de variantie van de schatter, namelijk $\mathrm{Var}[\widehat{\theta}]$, of soms ook gezien als de onzekerheid op de waarde van de schatter. De waarde van de variabele $\mathrm{Var}[\widehat{\theta}]$ of $\sigma^2_{\widehat{\theta}}$ hebben we in de vorige sectie ook geschat via de schatter $\widehat{\sigma_{\widehat{\theta}}}$.

Uiteindelijk willen we het resultaat van de steekproef van n metingen weergeven als

$$\widehat{\mu_\theta} \;=\; \widehat{\mu_x} \;=\; \overline{x} \pm \widehat{\sigma_{\widehat{\theta}}} \;=\; \overline{x} \pm \sqrt{\frac{s_x^2}{n}} \tag{6.29}$$

waar we één empirische waarde $\overline{x}$ hebben als schatting van $\widehat{\mu_\theta}$ van de locatie van de theoretische parameter θ_0 die de verwachtingswaarde μ_x van de stochastiek X is en één empirische waarde $\widehat{\sigma_{\widehat{\theta}}}$ voor de schatting van de onzekerheid of standaardafwijking op de gevonden waarde van $\widehat{\mu_\theta}$. Beide variabelen, $\widehat{\mu_\theta}$ en $\widehat{\sigma_{\widehat{\theta}}}$, zijn onafhankelijk.

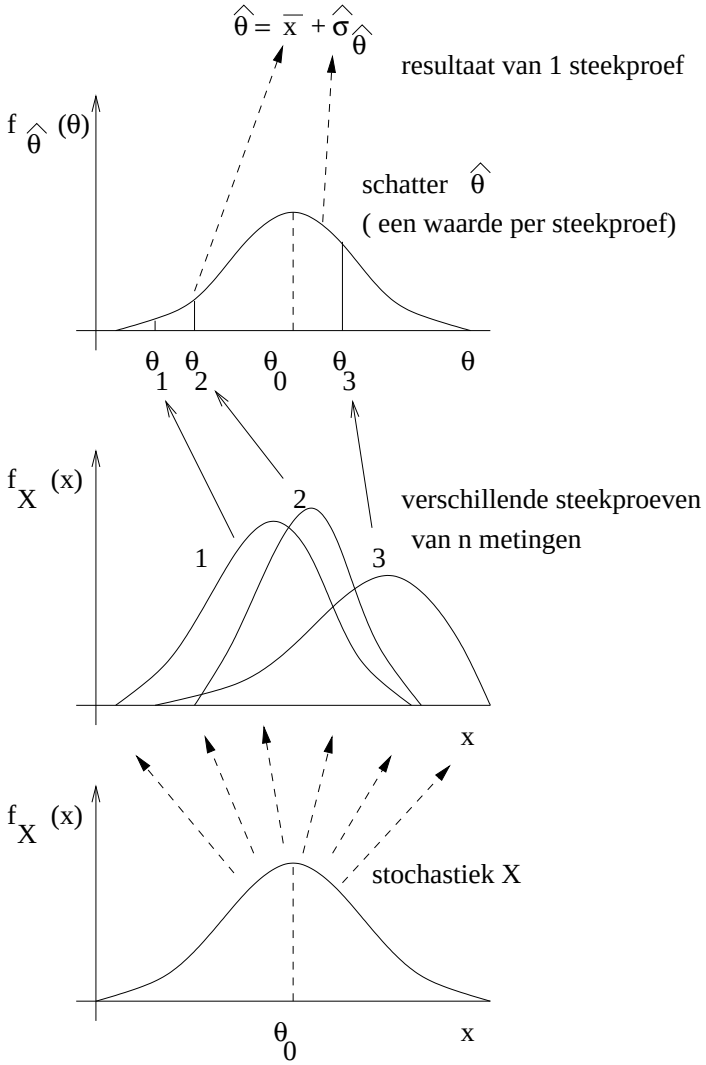
Ook hebben we een schatter besproken voor de variantie van de stochastiek X, namelijk

$$\widehat{\sigma_x^2} = s_x^2 \pm \sqrt{\frac{2}{n-1}} s_x^2 \tag{6.30}$$

die ook nuttig is om de onderliggende verdeling van stochastiek X te achterhalen.

In Figuur 6.2 worden alle begrippen nog eens samengevat in een illustratie. Het eindresultaat $\widehat{\mu_\theta} = \overline{x} \pm \widehat{\sigma_{\widehat{\theta}}}$ is bepaald door slechts één steekproef van n metingen. De waarde van de variabele $\widehat{\sigma_{\widehat{\theta}}}$ geeft een schatting van de spreiding van de verdeling $f_{\widehat{\theta}}(\theta)$, en is bekomen uit n empirische gegevens. De waarde van de grootheid $\overline{x}$ geeft de locatie weer van de steekproef, en dus niet van de constante waarde van de theoretische parameter θ_0. Via beide variabelen, $\overline{x}$ en $\widehat{\sigma_{\widehat{\theta}}}$ willen we iets interpreteren over de waarde van de theoretische parameter θ_0.

Omdat we slechts één steekproef van n metingen genomen hebben uit een zekere populatie, kunnen we niets zeggen over de onzekerheid op de theoretische waarde θ_0. We kunnen enkel iets zeggen over de onzekerheid op de waarde van de schatter $\widehat{\mu_\theta}$ en daaruit iets concluderen over de theoretische waarde θ_0. Merk op, dat er op de theoretische waarde θ_0 eigenlijk



Figuur 6.2: *Illustratie bij de begrippen rond schatters.*

geen onzekerheid bestaat omdat het een constante waarde is. Het heeft bijgevolg ook geen nut om na te denken over de onzekerheid op de theoretische waarde θ_0.

Veelal wordt de fout gemaakt om de steekproefwaarde van $\widehat{\theta}$, namelijk $\overline{x}$, te interpreteren als de waarde van de theoretische parameter θ_0 en wordt een normale waarschijnlijkheids-dichtheidsverdeling $N(\overline{x}, \sigma^2)$ rond die waarde geconstrueerd met standaardafwijking $\sigma = \widehat{\sigma_{\widehat{\theta}}}$ om aan te geven wat de meest waarschijnlijke waarde van θ_0 is. Zoals we gezien hebben in Figuur 6.2 is deze interpretatie verkeerd. De waarde van $\widehat{\theta}$, bekomen uit één enkele steekproef, is namelijk niet gelijk aan θ_0, maar heeft een spreiding van $\widehat{\sigma_{\widehat{\theta}}}$ rond θ_0.

Stel dat de schatter een normale verdeling $N(\mu_0, \sigma_{\widehat{\theta}})$ volgt, wat meestal het geval is, dan weten we uit Hoofdstuk 4 dat ongeveer 2/3 van de metingen van de stochastiek $\widehat{\theta}$ in het interval $[\mu_0 - \sigma_{\widehat{\theta}}, \mu_0 + \sigma_{\widehat{\theta}}]$ liggen

$$P\left(\mu_0 - \widehat{\sigma_{\widehat{\theta}}} \;\leq\; \overline{X} \;\leq\; \mu_0 + \widehat{\sigma_{\widehat{\theta}}}\right) \simeq 68\% \;\; . \tag{6.31}$$

Dit komt erop neer dat ongeveer 2/3 van de steekproeven binnen dit interval liggen. Wat kunnen we nu besluiten indien we slechts één steekproef uitvoeren?

Voor elke steekproef kunnen we ook een interval opstellen, namelijk $[\overline{x} - \widehat{\sigma_{\hat{\theta}}}, \overline{x} + \widehat{\sigma_{\hat{\theta}}}]$. We weten dat voor 1/3 van de steekproeven de waarde voor theoretische parameter, namelijk μ_0, buiten dit interval ligt. Indien we één willekeurige steekproef uitvoeren vinden we bijgevolg dat de theoretische waarde van de parameter μ_0 met een waarschijnlijkheid van 2/3 binnen het interval $[\overline{x} - \widehat{\sigma_{\hat{\theta}}}, \overline{x} + \widehat{\sigma_{\hat{\theta}}}]$ ligt of

$$P\left(\overline{x} - \widehat{\sigma_{\hat{\theta}}} \ \leq \ \mu_0 \ \leq \ \overline{x} + \widehat{\sigma_{\hat{\theta}}}\right) \simeq 68\% \ . \tag{6.32}$$

Niettegenstaande de bovenstaande redenering triviaal lijkt, vormt ze toch de basis van hevige discussies tussen statistici. De overgang van uitdrukking 6.31 naar uitdrukking 6.32 zorgt voor discussie. Volgens het Bayesiaanse kamp kan men enkel via de regel van Bayes overgaan van uitdrukking $f(\overline{x} \mid \mu_0)$ naar $f(\mu_0 \mid \overline{x})$. Maar laten we volgend voorbeeld eens nader bekijken. We wegen een leeg bord met een weegschaal en we verkrijgen een gewicht van 25.31 ± 0.14 gram. Stel nu dat we een beetje poeder op het bord leggen en opnieuwe wegen, we bekomen 25.51 ± 0.14 gram voor het totaal. Nu willen we weten wat het gewicht van de poeder is. Dit wordt gegeven door het verschil, namelijk 0.20 ± 0.20 gram [2]. Als we nu de waarschijnlijkheden uitrekenen en aannemen dat de variabelen een normale verdeling volgen, bekomen we een kans van ongeveer 16% dat het poeder een negatief gewicht heeft. Daar we nog geen anti-zwaartekracht poeder hebben, is dit uiteraard pure onzin. Het probleem komt echter van de regel van Bayes waarin we een uniforme *prior* of voorkennisfunctie hebben aangenomen. De tweede keer dat we de weegschaal gebruikt hebben, moesten we rekening houden met het feit dat we geen kleiner gewicht konden bekomen dan de eerste keer. Maar wie gaat natuurlijk zeggen welke *a priori* kennis juist is !!

Algemeen is onze redenering om tot uitdrukking 6.32 te komen, slechts geldig voor symmetrische verdelingen van de stochastiek X. Voor niet symmetrische verdelingen en indien n klein is, treden er afwijkingen op waarmee men moet rekening houden. Dit is nog steeds het onderwerp van verschillende internationale wetenschappelijke conferenties en het is bijgevolg niet de doelstelling van deze cursus om die allemaal uiteen te zetten. In Hoofdstuk 8 zullen we zogenaamde betrouwbaarheidsintervallen definiëren die informatie weergeven over θ_0, dit zonder aanname van *a priori* kennis. Deze werkwijze komt dan overeen met het ander kamp, namelijk de 'frequentisten'. In de meeste gevallen komen beide werkwijzen overeen !!

6.5 Voortplanting van verschillende types onzekerheden

Denken we nu terug aan het voorbeeld van het brandstofverbruik van een voertuig. Een maat R voor dit verbruik is de verhouding van de door het voertuig verbruikte brandstof (in liter) en de door het voertuig afgelegde afstand (in kilometer). Beide grootheden, aantal liter l en aantal kilometer k, worden beschreven door een stochastiek die een onbekende verdeling aanneemt. Bijgevolg kan men een schatter opstellen voor beide en een empirisch resultaat

[2] We zullen in de volgende sectie zien hoe we onzekerheden van twee metingen combineren.

bekomen $\widehat{\mu_l} = \overline{l} \pm \widehat{\sigma_{\overline{l}}}$ en $\widehat{\mu_k} = \overline{k} \pm \widehat{\sigma_{\overline{k}}}$. Hoe kunnen we nu van deze empirische gegevens overgaan naar een resultaat over R ?

We kunnen verschillende types onzekerheden onderscheiden bij het uitvoeren van een experiment of meting. Zo hebben we onzekerheden die niet aan bepaalde wetten voldoen, die men met andere woorden niet kan voorspellen. Hierin hebben we onder andere vergissingen, het verkeerd aflezen van een toestel, het verkeerd gebruik van een toestel, verkeerde handelingen, verkeerde berekeningen, enzovoort. Het is de betrachting van een wetenschapper om die 'blunders' te vermijden !! Een andere categorie bestaat uit onzekerheden die we wel kunnen voorspellen en bijgevolg bestuderen. We maken volgend onderscheid.

- **Systematische fouten**: Deze fouten hebben welbepaalde oorzaken die ook aan bepaalde wetten voldoen. Indien deze wetten gekend zijn, kan men de meting corrigeren of verbeteren.

- **Statistische onzekerheden**: Deze onzekerheden bekomen we door de fluctuaties in onze meetopstelling. Deze fluctuaties worden beschreven door een zekere stochastiek X via de verdeling $f_X(x \mid \theta_0)$ die afhankelijk is van de parameters die we willen meten. Door het groot aantal van dergelijke storende factoren die men ontmoet bij het uitvoeren van een experiment, volgt uit de centrale limietstelling dat de waargenomen verdeling voor stochastiek X een normale verdeling benadert.

Let op het feit dat we de eerste categorie 'fouten' noemen, terwijl we bij de tweede categorie spreken van 'onzekerheden'. Dit maakt een groot verschil voor de interpretatie die meestal verkeerd gebruikt wordt. Een fout resulteert in een systematische verschuiving van het resultaat. Deze fout zal een verschuiving introduceren die al dan niet gekend is, maar heeft één unieke waarde. Een fout kan bijvoorbeeld een systematische invloed zijn van een weegschaal die niet goed geijkt is. Een weegschaal kan systematisch 10 gram te weinig wegen. Indien we met dergelijke weegschaal een meting doen, die overigens een echte nauwkeurigheid van 1 gram heeft, zullen we denken dat we heel precies wegen. Een belangrijk onderdeel van een experiment is bijgevolg de voorbereiding. We moeten voordien heel nauwkeurig alle systematische fouten wegwerken. Een toestel wordt geijkt met een welbepaalde nauwkeurigheid, deze nauwkeurigheid is meestal gegeven en kunnen we beschouwen als een schatting van de grootte van de systematische fout. Indien men zegt dat de nauwkeurigheid van het toestel 2% is, bedoelt men dat men er bijna zeker van is dat de echte systematische fout kleiner is dan 2%. Men maakt namelijk ook slechts een statistische schatting van een systematische fout. Een systematische fout kunnen we bijgevolg niet wegwerken door simpelweg meer gegevens te nemen.

Een onzekerheid wordt veroorzaakt door een stochastisch proces en mag men niet interpreteren als een fout. We hebben namelijk geen fout gemaakt in ons experiment, maar we zijn afhankelijk van willekeurige fluctuaties van het proces onder studie die een zekere verdeling volgen. Indien we met een perfect geijkte chronometer de tijd meten dat een knikker nodig heeft om van hoogte A naar hoogte $A - 100$ te vallen, bekomen we een onzekerheid. Zo zijn de observatiesnelheid van onze ogen en de reactiesnelheid van onze handen niet oneindig, en moeten we rekening houden met fluctuaties. Zoals we gezien hebben in dit hoofdstuk, kunnen we deze onzekerheid verkleinen door meer gegevens te nemen en van

al deze gegevens het gemiddelde te gebruiken. De onzekerheid op het gemiddelde wordt kleiner met een factor $1/\sqrt{n}$ waar n het aantal metingen is.

Voor het opstellen van de formules voor de voortplanting van onzekerheden kunnen we in het algemeen een functie $z = f(x_1, x_2, ..., x_k)$ beschouwen. Laat ons voor de eenvoud veronderstellen dat de stochastische variabelen X_j allen onafhankelijk zijn. Zoals in de meeste gevallen kunnen we aannemen dat z een langzaam variërende functie is, zodat deze kan benaderd worden door haar raakvlak in het gebied waar de waarschijnlijkheidsdichtheid groot is. Dit gebied ligt bijgevolg rond de verwachtingswaarde $\vec{\mu_j}$, die een punt voorstelt in de k-dimensionale steekproefruimte, en heeft een spreiding van σ_j die verschillend kan zijn in elke richting j. Met deze voorwaarde kunnen we de functie benaderen met een Taylor ontwikkeling, namelijk

$$z = f(x_1, x_2, ..., x_k) \simeq f(\mu_1, \mu_2, ..., \mu_k) + \sum_{j=1}^{k} \frac{\partial f}{\partial x_j}(x_j - \mu_j) + ... \tag{6.33}$$

waar we de partiële afgeleiden nemen in het punt $(\mu_1, \mu_2, ..., \mu_k)$ en omdat de functie niet te snel varieert, kunnen we alle hogere orde termen verwaarlozen. Indien de vergelijking $z = f(x_1, x_2, ..., x_k)$ een lineair verband weergeeft tussen de variabelen, dan is de benadering 6.33 exact. Dit omdat de hogere orde afgeleiden toch nul zijn. We vinden dat z een lineaire combinatie is van stochastische variabelen.

We vinden de volgende uitdrukking voor de verwachtingswaarde $E[Z]$ van de stochastiek Z

$$E[Z] \simeq f(\mu_1, \mu_2, ..., \mu_k) \tag{6.34}$$

en voor de variantie $\mathrm{Var}[Z]$ vinden we

$$\mathrm{Var}[Z] \simeq \sum_{j=1}^{k} \left(\frac{\partial f}{\partial x_j}\right)^2 \sigma_{x_j}^2 \quad . \tag{6.35}$$

Dit verkrijgen we omdat de variantie van een constante gelijk is aan nul en door gebruik te maken van de andere eigenschappen van de variantie besproken in Hoofdstuk 3. Als de veranderlijken X_j afhankelijk zijn, dan geldt deze uitdrukking niet en moeten we de correlatiecoëfficiënten in rekening brengen. Deze formules kunnen we veel eenvoudiger neerschrijven in matrixnotatie en zullen besproken worden in de hogere studiejaren.

De vergelijking 6.35 is essentieel bij het uitvoeren van een experiment waar men een verband heeft tussen k grootheden $\{x_j \mid j \in \{1, 2, ..., k\}\}$ die allen behalve één, namelijk grootheid g, empirisch gemeten zijn. De verschillende resultaten van de $(k-1)$ metingen kunnen we samenvatten als

$$\widehat{\mu_{x_j}} = \overline{x_j} \pm \sqrt{\frac{s_{x_j}^2}{n_j}} \qquad \text{voor } j \neq g \tag{6.36}$$

waar iedere schatter $\widehat{\mu_{x_j}}$ $(j \neq g)$ bepaald wordt uit in totaal n_i $(j \neq g)$ metingen. Met behulp van vergelijkingen 6.34 en 6.35 kunnen we de waarde van en de onzekerheid op $\widehat{\mu_{x_g}}$ bepalen.

Neem een eenvoudige vergelijking als voorbeeld

$$z = f(x_1, x_2, x_3) = a_1 + a_2 x_1 + a_3 x_2^2 + \frac{a_4}{x_3} + a_5 x_1 x_2 \tag{6.37}$$

waar $\{a_l \mid l \in \{1, 2, 3, 4, 5\}\}$ constante getallen zijn. Omdat die vergelijking dus niet lineair is, moeten we via de Taylor ontwikkeling een benadering maken van de vergelijking, met andere woorden de vergelijking lineariseren. De vergelijking voor de voortplanting van onzekerheden 6.35 geldt bijgevolg maar als een benadering. We bekomen volgende empirische metingen [3] van de stochastische variabelen x_1, x_2 en x_3

$$\widehat{\mu_{x_1}} = -3.0 \pm 0.4 \tag{6.38}$$

$$\widehat{\mu_{x_2}} = +5.0 \pm 0.2 \tag{6.39}$$

$$\widehat{\mu_{x_3}} = +0.5 \pm 0.5 \tag{6.40}$$

Wat kunnen we nu concluderen over $\widehat{\mu_z}$? Met de vergelijking 6.35 bekomen we

$$\sigma_z^2 \simeq (a_2 + a_5 x_2)^2 \sigma_{x_1}^2 + (2a_3 x_2 + a_5 x_1)^2 \sigma_{x_2}^2 + \left(a_4 \frac{1}{x_3^2}\right)^2 \sigma_{x_3}^2 \ . \tag{6.41}$$

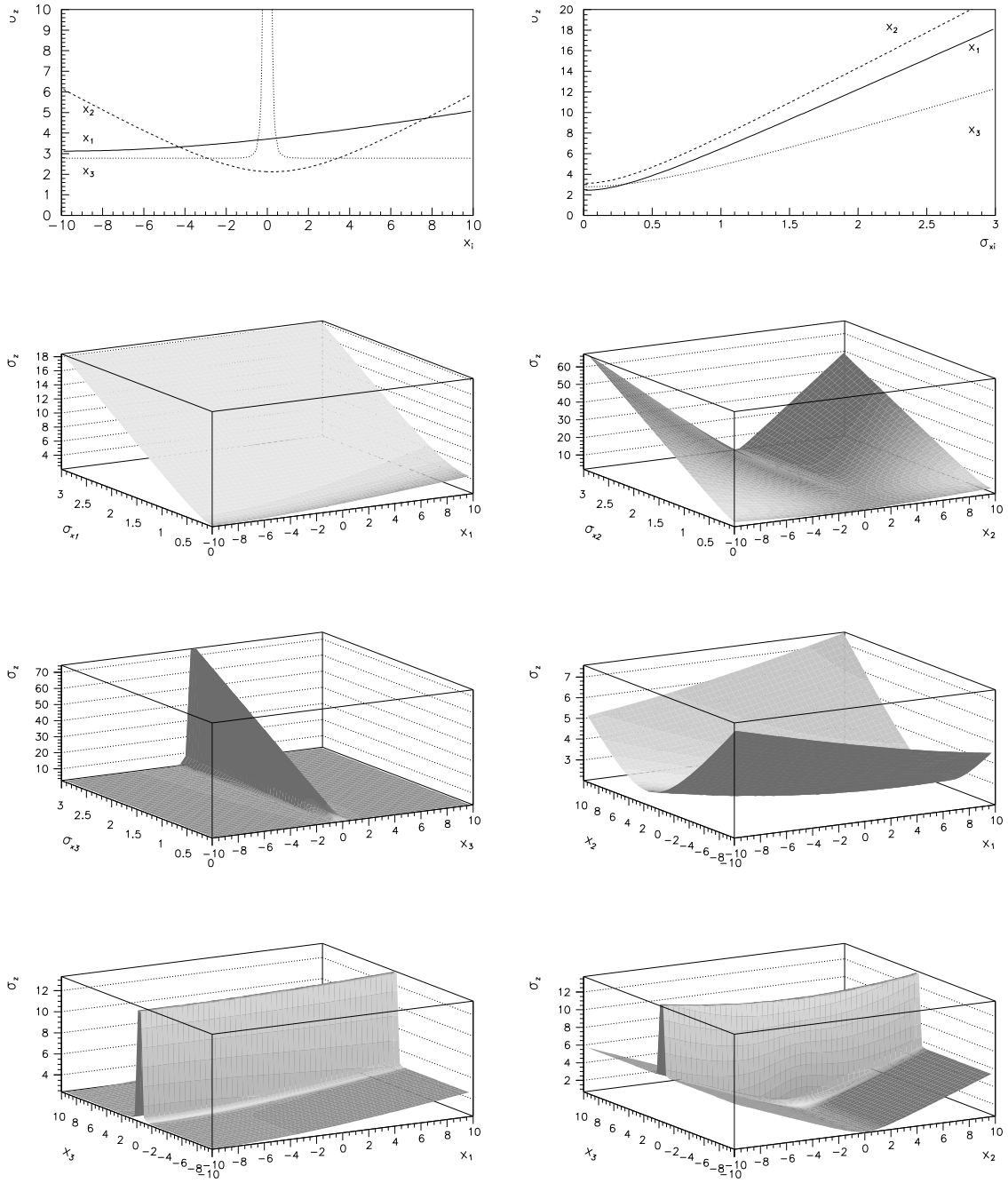
Stel de numerieke waarden voor a_l gelijk aan 1 voor alle l, dan krijgen we voor de variantie van de stochastische variabele Z de waarde $\sigma_z^2 = 11.7$ en bijgevolg het resultaat

$$\widehat{\mu_z} = 10.0 \pm 3.4 \ . \tag{6.42}$$

In Figuur 6.3 zien we hoe de waarde van de onzekerheid σ_z verandert ten opzichte van de variabelen in vergelijking 6.41. Duidelijk zien we dat, indien de waarde van x_3 nul benadert, de onzekerheid σ_z enorm toeneemt, wat ook te verwachten was. Indien de onzekerheden σ_{x_j} toenemen, zal ook de onzekerheid σ_z toenemen. Dit is alweer een logisch gevolg, wat een goede methode kan zijn om eventuele fouten bij de berekening op te sporen. In de vergelijking 6.37 zien we een term die de stochastieken X_1 en X_2 combineert. Door deze term is de invloed van de waarde van x_1 en x_2 op σ_z gecorreleerd, zoals we zien in de figuur. Dit is niet zo voor de combinaties $x_1 \leftrightarrow x_3$ en $x_2 \leftrightarrow x_3$.

Het bestuderen van dergelijke grafieken is soms belangrijk bij het optimaliseren van een experiment. Indien we de waarde van Z zo nauwkeurig mogelijk willen bepalen, kunnen we met behulp van deze grafieken nagaan welke variabele X_j we het nauwkeurigst moeten meten. Voor dit voorbeeld heeft de waarde van σ_{x_2} de grootste invloed op σ_z. Maar deze conclusie kan in sommige gevallen afhankelijk zijn van de waarde van de schatters voor μ_{X_j}.

[3] Laat ons gemakshalve de dimensies vergeten.



Figuur 6.3: *Afhankelijkheid van de onzekerheid σ_z ten opzichte van de variabelen in de vergelijking.*

Hoofdstuk 7

De methode van de kleinste kwadraten

"I never met a man so ignorant that I couldn't learn something from him.''

Galileo Galilei (1564-1642)

In vorig hoofdstuk hebben we de eigenschappen van een goede schatter besproken. Er bestaan verschillende methoden om de waarde van de parameters van een theoretische verdeling te schatten. Eén daarvan bespreken we hier, namelijk de methode van de kleinste kwadraten of de zogenaamde *least square* methode. Deze wordt soms ook de chi-kwadraat methode genoemd. De methode werd voor het eerst gebruikt in 1805 door Adrien-Marie Legendre (1752-1833) terwijl Gauss beweert de methode reeds te gebruiken in 1795 voor het bepalen van de beweging van hemellichamen.

7.1 Schatten van de verwachtingswaarde

Beschouw een verzameling van n empirische metingen van stochastiek X. De methode van de kleinste kwadraten bestaat erin, de theoretische waarde μ_0 voor de verwachtingswaarde $E[X]$ te schatten met schatter $\widehat{\mu_x}$. De waarde μ_x van de schatter die de waarde van Q^2 of χ^2 minimaliseert

$$Q^2(\widehat{\mu_x}) = \chi^2(\widehat{\mu_x}) = \sum_{i=1}^{n} \left(\frac{x_i - \widehat{\mu_x}}{\sigma_{x_i}} \right)^2 \tag{7.1}$$

is de beste schatting voor μ_0. Let erop, door deze relatie tussen de stochastieken X_i, volgt deze grootheid een χ^2_{n-1}-verdeling en niet een χ^2_n-verdeling. Een functie minimaliseren, doet men door zijn afgeleide gelijk te stellen aan nul en bijgevolg

$$\left. \frac{\partial \chi^2(\mu_x)}{\partial \mu_x} \right|_{\mu_x = \widehat{\mu_x}} = -2 \sum_{i=1}^{n} \frac{x_i - \widehat{\mu_x}}{\sigma_{x_i}^2} = 0 \ . \tag{7.2}$$

Deze lineaire vergelijking kunnen we oplossen naar $\widehat{\mu_x}$

$$\widehat{\mu_x} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_{x_i}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_{x_i}^2}} \quad . \tag{7.3}$$

De variantie van de verdeling van $\widehat{\mu_x}$ kunnen we bepalen met behulp van de kennis van de varianties van de stochastieken X_i, namelijk

$$\mathrm{Var}[\widehat{\mu_x}] = \sum_{i=1}^{n} \left(\frac{\partial \widehat{\mu_x}}{\partial x_i} \right)^2 \mathrm{Var}[X_i] = \left(\frac{1}{\sum_{i=1}^{n} \left(\frac{1}{\sigma_{x_i}} \right)^2} \right)^2 \sum_{i=1}^{n} \frac{\mathrm{Var}[X_i]}{\sigma_{x_i}^4} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_{x_i}^2}} \quad . \tag{7.4}$$

Hiermee tonen we een mooie eigenschap aan van de kleinste kwadraten schatter, namelijk dat bij het toevoegen van een meting, de onzekerheid $\mathrm{Var}[\widehat{\mu_x}]$ nooit groter wordt. De schatter is bijgevolg efficiënt. Indien we de verwachtingswaarde van de onderliggende theoretische verdeling schatten, vinden we dat de variantie onafhankelijk is van de waarde Q^2 of χ^2. De variantie $\mathrm{Var}[\widehat{\mu_x}]$ is echter wel afhankelijk van de vorm van de verdeling $\chi^2(\mu_x)$ daar

$$\left. \frac{\partial^2 \chi^2(\mu_x)}{\partial \mu_x^2} \right|_{\mu_x = \widehat{\mu_x}} = 2 \sum_{i=1}^{n} \frac{1}{\sigma_{x_i}^2} = \frac{2}{\mathrm{Var}[\widehat{\mu_x}]} \quad . \tag{7.5}$$

Merk op dat alle hogere afgeleiden van $\chi^2(\mu_x)$ gelijk zijn aan nul. Hieruit kunnen we besluiten dat $\chi^2(\mu_x)$ een parabolische functie is van μ_x. De Taylor ontwikkeling is

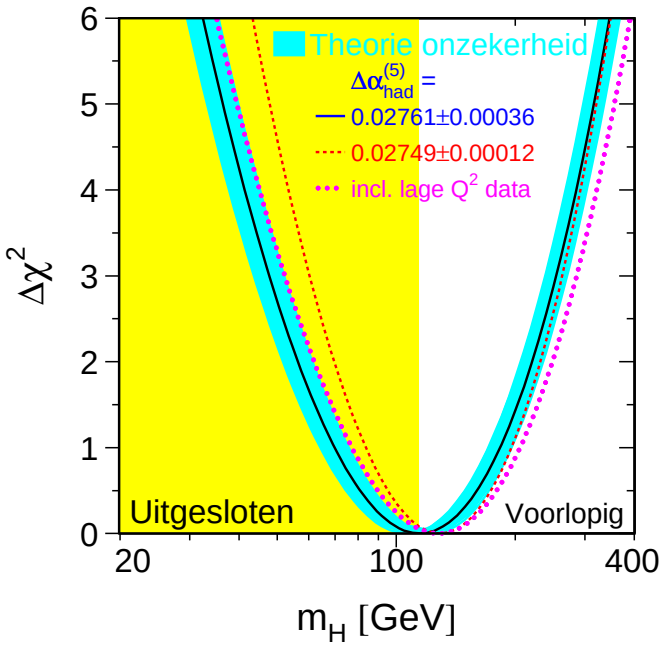
$$\chi^2(\mu_x) = \chi^2(\widehat{\mu_x}) + \frac{(\widehat{\mu_x} - \mu_x)^2}{\mathrm{Var}[\widehat{\mu_x}]} \quad . \tag{7.6}$$

We kunnen de onzekerheid op $\widehat{\mu_x}$ vinden door een waarde μ_x te zoeken waarvoor $\Delta\chi^2(\mu_x) = \chi^2(\mu_x) - \chi^2(\widehat{\mu_x})$ een zekere waarde aanneemt. Met behulp van de vergelijking 7.6 zien we dat $\Delta\chi^2(\mu_x) = 1$ voorkomt indien $(\widehat{\mu_x} - \mu_x)^2 = \mathrm{Var}[\widehat{\mu_x}]$. Bijgevolg, indien μ_x één standaardafwijking afwijkt van $\widehat{\mu_x}$, heeft $\Delta\chi^2(\mu_x)$ een waarde gelijk aan 1. Algemeen geldt dat voor waarden van μ_x waarvoor

$$\Delta\chi^2(\mu_x) = \chi^2(\mu_x) - \chi^2(\widehat{\mu_x}) = n^2 \tag{7.7}$$

deze n standaardafwijkingen verwijderd liggen van de beste waarde van de schatter $\widehat{\mu_x}$.

In Figuur 7.1 vinden we een realistisch voorbeeld van een $\Delta\chi^2(m)$ functie. Deze curve is het resultaat van een schatting van de massa m van het nog niet ontdekte Brout-Englert-Higgs deeltje. Dit deeltje is een essentiële bouwsteen binnen het Standaard Model van de elementaire deeltjes en is het enige nog niet ontdekte elementair deeltje dat voorspeld wordt door het Standaard Model. Er zijn echter vele andere metingen uitgevoerd van parameters van het Standaard Model en deze staan via theoretische vergelijkingen in verband met de massa van het Brout-Englert-Higgs deeltje. Met de kleinste kwadraten methode kunnen we deze empirische gegevens gebruiken om een schatting te maken van zijn massa. Het resultaat wordt weergegeven met een $\Delta\chi^2$ curve in de figuur. De beste waarde van de schatter $\widehat{m}$ ligt bij het minimum van de $\Delta\chi^2(m)$ curve en de waarden voor $\widehat{m} \pm \widehat{\sigma_m}$ vinden we waar $\Delta\chi^2(m)$ gelijk is aan $\Delta\chi^2(\widehat{m}) \pm 1$. Op deze manier kunnen we een schatting $\widehat{\sigma_m}$ maken van de



Figuur 7.1: *Voorbeeld van een $\Delta\chi^2(m)$ curve voor de massa van het nog niet ontdekte Brout-Englert-Higgs deeltje.*

standaardafwijking van de schatter $\widehat{m}$. De interpretatie van deze standaardafwijkingen loopt gelijk met deze besproken in Hoofdstuk 6. Deze empirische gegevens komen uit slechts één enkele steekproef. Indien we m steekproeven uitvoeren, kunnen we m keer het interval $[\widehat{m} - \widehat{\sigma_{\widehat{m}}}, \widehat{m} + \widehat{\sigma_{\widehat{m}}}]$ opstellen. De echte theoretische waarde m_0 van de massa zal in ongeveer 68% van de steekproefintervallen liggen. Dit uiteraard indien het Standaard Model de juiste verbanden heeft voorspeld tussen zijn parameters.

7.2 Lineair verband

Hierboven hebben we de kleinste kwadraten methode opgesteld voor een aantal metingen n van één en dezelfde grootheid, beschreven door stochastiek X. Dit willen we nu veralgemenen naar een aantal metingen n die de waarde y_i opleveren van grootheid Y welke nu wel afhankelijk is van een andere grootheid X die niet noodzakelijk een stochastische verdeling moet volgen. Denk aan het brandstofverbruik van een auto. We kunnen het aantal verbruikte liters brandstof meten na een vooraf vastgelegde afstand. Met andere woorden meten we de grootheid Y voor n verschillende waarden van X, namelijk $\{x_i \mid i \in \{1, 2, ..., n\}\}$. Om de methode te vereenvoudigen, stellen we dat alle waarden x_i exact gekend zijn en bijgevolg geen onzekerheid hebben. We kunnen dit ook benaderen door aan te nemen dat de bepaling van grootheid X een veel kleinere onzekerheid heeft, vergeleken met de onzekerheid op

de meting van grootheid Y, of $\sigma_x^2 \ll \sigma_y^2$. Voor elke x_i meten we een waarde van y_i voor grootheid Y met een onzekerheid $\sigma_{y_i}^2$. Om alweer de methode te vereenvoudigen moeten we aannemen dan $\sigma_{y_i}^2$ niet afhankelijk is van y_i.

Beschouw volgend lineair model voor het verband $y = f(x)$

$$y(x) = \theta_1 h_1(x) + \theta_2 h_2(x) + ... + \theta_k h_k(x) \tag{7.8}$$

waar we k parameters $\{\theta_j \mid j \in \{1, 2, ..., n\}\}$ introduceren die onbekende coëfficiënten zijn in de functie. We willen nu een kleinste kwadraten methode [1] opstellen die met de n empirische gegevens $\{x_i, y_i, \sigma_{y_i}\}$ de beste waarden voor de coëfficiënten θ_j bepaalt. De theoretische waarde van de parameters of coëfficiënten van vergelijking 7.8 zijn constanten en uiteraard ongekend. We noteren ze met $\theta_{j,0}$. Indien we geen verschuivingen willen in de resultaten voor de geschatte parameters $\widehat{\theta}_j$, of $b_j(n) = \theta_{j,0} - E[\widehat{\theta}_j] = 0$, moeten we de k functies $h_j(x)$ exact kennen. Deze functies $h_j(x)$ moeten een eenduidige waarde weergeven indien men ze berekent voor een willekeurige waarde van grootheid X. Indien men alle parameters θ_j wil schatten met schatters $\widehat{\theta}_j$ moeten ook alle functies $h_j(x)$ onafhankelijk zijn. Dit impliceert dat geen enkele $h_j(x)$ kan uitgedrukt worden als een lineaire combinatie van de andere $k - 1$ functies. Is dit wel zo, dan kan men deze termen samennemen en is ook de parameter θ_j afhankelijk van de andere $k - 1$ parameters en zal hij bijgevolg niet onafhankelijk geschat kunnen worden. Er zijn geen andere voorwaarden voor de functies $h_j(x)$. Het begrip van een 'lineair' model voor de kleinste kwadraten methode slaat op de parameters θ_j en niet op de functies $h_j(x)$. Indien de vergelijking 7.8 niet lineair is voor alle θ_j, dan kan men die lineariseren door slechts de eerste termen mee te nemen van een Taylor ontwikkeling van $y(x)$.

Het schatten van de parameters θ_j van het theoretisch of analytisch verband $y = f(x)$ tussen grootheden X en Y, gaat ervan uit dat de stochastische afwijkingen ϵ_j van de empirische gegevens ten opzichte van de theoretische curve te wijten zijn aan de onzekerheid op de meting van grootheid Y. Deze afwijkingen ϵ_j moeten niet noodzakelijk een normale verdeling volgen. We kunnen dit noteren voor elke meting $i \in \{1, 2, ..., n\}$ als

$$y_i = y(x_i) + \epsilon_i = \sum_{j=1}^{k} \theta_j h_j(x_i) + \epsilon_i \tag{7.9}$$

waar de ongekende verschuiving ϵ_i op y_i de volgende eigenschappen heeft

$$E[\epsilon_i] = 0 \qquad \text{en} \qquad \text{Var}[\epsilon_i] = \sigma_{y_i}^2 \tag{7.10}$$

en waar $\sigma_{y_i}^2$ gekend is met behulp van een schatting in een vorig experiment. Voor de eenvoud veronderstellen we dat alle metingen y_i onafhankelijk zijn, anders moeten we overgaan op matrixnotatie om de correlaties in rekening te brengen. De waarden x_i mogen willekeurig gekozen worden, en mogen zelfs dezelfde waarden aannemen. We zullen echter zien dat uit de n metingen er minstens k verschillende waarden x_i nodig zijn om de waarde van k parameters θ_j te schatten.

[1] In de literatuur wordt deze kleinste kwadraten methode om parameters van een functie te schatten of te 'fi tten' ook wel ten onrechte een regressie-analyse genoemd. Beide concepten hebben echter een verschillende betekenis.

Voor deze probleemstelling kunnen we Q^2 of χ^2 definiëren als

$$Q^2 = \chi^2 = \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_{y_i}^2} \tag{7.11}$$

of

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - y(x_i)}{\sigma_{y_i}} \right)^2 = \sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2} \left(y_i - \sum_{j=1}^{k} \theta_j h_j(x_i) \right)^2 \quad . \tag{7.12}$$

Deze χ^2 volgt een χ_n^2-verdeling enkel indien de verschuivingen ϵ_i een normale verdeling volgt. Vandaar dat men soms de notatie Q^2 gebruikt in plaats van χ^2. Sinds de waarde $\theta_{j,0}$ van de exacte theoretische parameters θ_j niet gekend is, kunnen we ook de exacte waarde van χ^2 niet bepalen. De kleinste kwadraten methode schat de waarde van de parameters met een schatter $\widehat{\theta_j}$. De waarde van $\widehat{\theta_j}$ die χ^2 minimaliseert, stelt men als de beste schatter en wordt gevonden door de afgeleide van χ^2 naar de parameters θ_j gelijk te stellen aan nul

$$\frac{\partial \chi^2(\vec{\theta})}{\partial \theta_l} = 2 \sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2} \left(y_i - \sum_{j=1}^{k} \widehat{\theta_j} h_j(x_i) \right) (-h_l(x_i)) = 0 \quad . \tag{7.13}$$

Dit levert een stelsel van k lineaire vergelijkingen met k onbekende parameters $\widehat{\theta_j}$. In de Algebra cursussen zullen jullie leren dergelijke stelsels op te lossen om de waarden van de parameters $\widehat{\theta_j}$ te bekomen. De onzekerheid op deze parameters bepalen we door alweer de $\Delta\chi^2(\vec{\theta})$ curve op te stellen. We kunnen alle parameters θ_j variëren en de k-dimensionale ruimte afgaan om de punten $\vec{\theta}$ te vinden waar

$$\Delta\chi^2(\vec{\theta}) = \Delta\chi^2(\vec{\widehat{\theta}}) + 1 \quad . \tag{7.14}$$

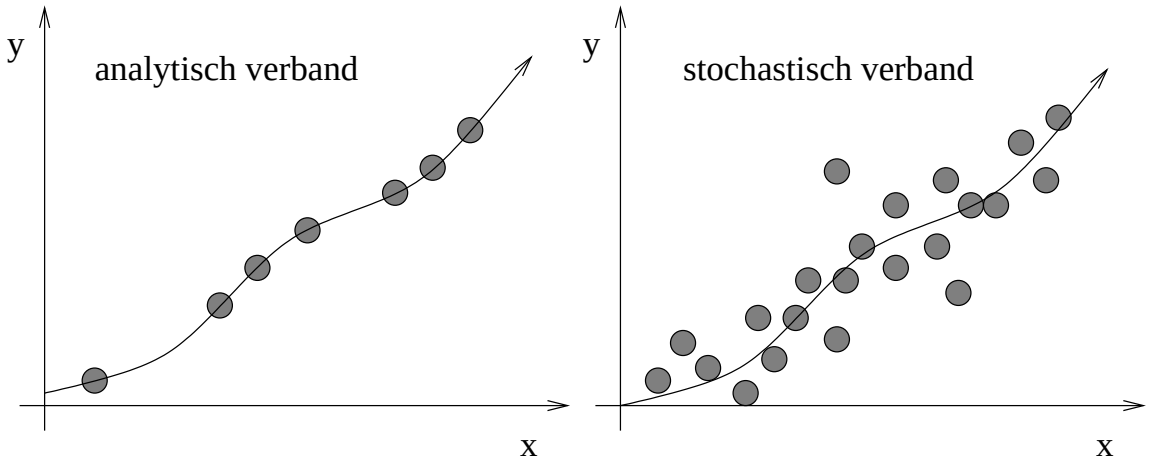
De (k-1)-dimensionale deelruimte van de parameterruimte $\vec{\theta}$ waar deze vergelijking geldt, definiëren we als de 68%-contour. De interpretatie is alweer identiek. Indien we m steekproeven uitvoeren, zal voor 68% van de steekproeven het geschatte interval de exacte theoretische waarde $\vec{\theta_0}$ bevatten.

Alle bovenstaande utdrukkingen en begrippen zijn eenvoudig te veralgemenen naar meerdere dimensies voor de stochastiek X, zodat $y = f(\vec{x})$.

In het geval dat we héél veel ($n \to \infty$) empirische gegevens hebben, kunnen we de coëfficiënten θ_j bepalen met een héél kleine onzekerheid. Met deze informatie zou men denken dat we heel precies kunnen voorspellen welke waarde y_{n+1} we zullen uitkomen indien x_{n+1} gekend is. Dit is helaas niet zo, daar we het nog steeds hebben over een stochastisch proces van de meting van grootheid Y. Er is een belangrijk verschil tussen een analytisch verband en een stochastisch verband, zie Figuur 7.2.

In de bovenstaande afleiding van de kleinste kwadraten methode hebben we veel veronderstellingen gemaakt. Er bestaan echter verschillende alternatieve methoden die elk van deze, soms minder practische voorwaarden, niet nodig hebben. Het bespreken van deze lange lijst behoort niet tot de doelstellingen van deze cursus.

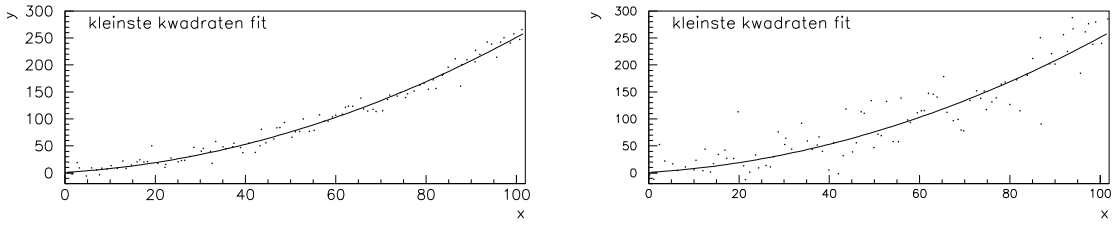
Figuur 7.3 illustreert een kleinste kwadraten fit voor twee verzameling van n gegevens. Beiden volgen een theoretische curve



Figuur 7.2: *Deze twee grafieken illustreren het verschil tussen een analytisch en een stochastisch verband.*

$$y(x) = \theta_1 + \theta_2 x + \theta_3 x^2 \tag{7.15}$$

met drie parameters $\theta_1 = 1$, $\theta_2 = 0.5$ en $\theta_3 = 0.02$. Deze theoretische curve wordt weergegeven in de grafiek met de exacte theoretische parameters. Het verschil tussen beide verzamelingen is de onzekerheid op de meting van de grootheid Y, namelijk σ_y. In de rechtse grafiek is de onzekerheid σ_y drie keer kleiner dan in de linkse grafiek. Bijgevolg zal de onzekerheid op de geschatte parameters θ_j kleiner zijn indien we de verzameling empirische gegevens van de rechtse grafiek gebruiken.



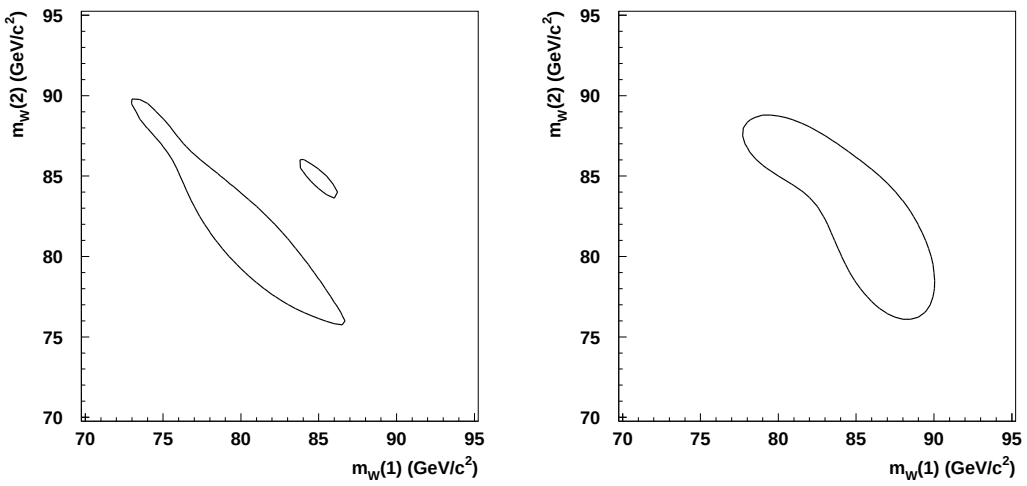
Figuur 7.3: *Illustratie van een kleinste kwadraten fit voor twee verschillende verzamelingen gegevens die elk dezelfde theoretische verdeling volgen, maar waar σ_y verschillend is.*

Een eerste toepassing van de kleinste kwadraten methode werd reeds gebruikt door Legendre en Gauss bij het bestuderen van de beweging van hemellichamen, maar dan onder een andere vorm. Stel dat we de massa van een hemellichaam splitsen in heel kleine deeltjes die onderling gelijk zijn en klein genoeg zijn om als punten beschouwd te worden. De som van de kwadraten van de afstanden tussen de puntdeeltjes en het centrum van de zwaartekracht, het zwaartepunt, is mimimaal. De som van de kwadraten van de relatieve afstanden tussen de puntdeeltjes en een willekeurig punt in de ruimte, verschillend van het zwaartepunt, zal

steeds groter zijn. Het zwaartepunt minimaliseert de som van de kwadraten en men bereikt bijgevolg de kleinste kwadraten.

7.3 Niet-lineair verband

Indien we een niet-lineair verband hebben tussen de stochastische grootheden X en Y, kunnen we de vergelijkingen $\frac{\partial \chi^2(\vec{\theta})}{\partial \theta_j}$ niet analytisch oplossen. We moeten met de computer iteratief de waarden voor $\vec{\theta}$ zoeken die de waarde van $\chi^2(\vec{\theta})$ minimaliseert. Bij het uitvoeren van dergelijke minimalisatie kan je op verschillende moeilijkheden stuiten. De $\Delta\chi^2(\vec{\theta})$ curve kan namelijk verschillende lokale minima hebben en bijgevolg kan de 68%-contour bestaan uit verschillende afgescheiden deelruimten. Indien we de minimalisatie in de verkeerde deelruimte starten, hebben de computerprogramma's meestal problemen om het globale minimum te vinden. Ze convergeren naar het lokale minimum in de nabijheid van de startwaarde van $\vec{\theta}$, en bereiken nooit het globale minimum van de $\Delta\chi^2(\vec{\theta})$ curve. Meestal kunnen het verband $y(x)$ in de kleinste kwadraten methode lineariseren om toch exacte analytische resultaten te bekomen.



Figuur 7.4: *Twee voorbeelden van een 68%-contour voor een schatting van twee parameters, namelijk m_W^1 en m_W^2.*

In Figuur 7.4 vinden we twee voorbeelden van een 68%-contour bepaald uit een kleinste kwadraten 'fit' voor twee parameters, m_W^1 en m_W^2. De LEP versneller in CERN nabij Génève heeft tussen de jaren 1990 en 2000 elektronen (e^-) en positronen (e^+) versneld. De DELPHI deeltjesdetector heeft botsingen van deze twee deeltjes waargenomen. In deze botsingen kwamen gebeurtenissen voor waar twee W bosonen werden geproduceerd die zelf vervallen in vele secundaire deeltjes. Indien men zo'n gebeurtenis bestudeert, meten we héél veel verschillende grootheden $\vec{X}$ die men in verband kan brengen met de massa's van beide W bosonen (ongeveer 80 GeV/c^2). Deze massa's, m_W^1 en m_W^2, beschouwen we als parameters in een lineaire vergelijking analoog aan 7.8 en kunnen we schatten met de kleinste

kwadraten methode. De resultaten kunnen we grafisch weergeven als 68%-contouren. De figuur illustreert dergelijke contouren voor twee van deze gebeurtenissen. We bemerken dat een 68%-contour bijzondere vormen kan aannemen en dat deze soms twee of meerdere afzonderlijke gebieden in de parameter ruimte kan omsluiten. Bij het oplossen van dergelijk complex probleem stuit met op verschillende niet-lineaire effecten. Deze hebben we hier vereenvoudigd door het totale resultaat op te splitsen in een superpositie van verschillende lineaire problemen. De som van dergelijke twee-dimensionale parabolische $\Delta\chi^2(m_W^1, m_W^2)$ curven, is opnieuw een functie met verschillende lokale minima.

7.4 Toepassing : Bepalen van de beste rechte

Een belangrijke toepassing van de kleinste kwadraten methode is het bepalen van het lineair verband tussen twee grootheden of het fitten van de beste rechte. Beschouw het volgende lineair verband tussen stochastieken X en Y die beiden n keer gemeten worden

$$y = a_0 x + b_0 \quad . \tag{7.16}$$

Met behulp van de n metingen $\{(y_i, x_i, \sigma_{y_i}) \mid i \in \{1, 2, ..., n\}\}$ willen we schatters $\widehat{a}$ en $\widehat{b}$ opstellen voor de coëfficiënten van de lineaire vergelijking. We werken in de veronderstelling dan de onzekerheid op de grootheid X verwaarloosbaar is, of $\sigma_{x_i} \ll \sigma_{y_i}$. Met de kleinste kwadraten methode kunnen we de waarde van de twee schatters bekomen door de Q^2 of χ^2 te minimaliseren. We bekomen

$$Q^2 = \chi^2 = \sum_{i=1}^{n} \left(\frac{\epsilon_i}{\sigma_{y_i}}\right)^2 = \sum_{i=1}^{n} \left[\frac{1}{\sigma_{y_i}^2}(y_i - \widehat{a}x_i - \widehat{b})^2\right] \tag{7.17}$$

die we moeten minimaliseren via de afgeleiden

$$\frac{\partial\chi^2(\widehat{a}, \widehat{b})}{\partial\widehat{a}} = 0 \tag{7.18}$$

en

$$\frac{\partial\chi^2(\widehat{a}, \widehat{b})}{\partial\widehat{b}} = 0 \quad . \tag{7.19}$$

Dit is een stelsel van twee vergelijkingen met twee onbekenden. Als we de vergelijkingen uitwerken, bekomen we

$$\left\{\begin{array}{rcccccl} \frac{\partial\chi^2(\widehat{a},\widehat{b})}{\partial\widehat{a}} & = & \frac{\partial}{\partial\widehat{a}}\sum_{i=1}^{n}\left[\frac{1}{\sigma_{y_i}^2}(y_i - \widehat{a}x_i - \widehat{b})^2\right] & = & -2\sum_{i=1}^{n}\frac{x_i}{\sigma_{y_i}^2}(y_i - \widehat{a}x_i - \widehat{b}) & = & 0 \\ \frac{\partial\chi^2(\widehat{a},\widehat{b})}{\partial\widehat{b}} & = & \frac{\partial}{\partial\widehat{b}}\sum_{i=1}^{n}\left[\frac{1}{\sigma_{y_i}^2}(y_i - \widehat{a}x_i - \widehat{b})^2\right] & = & -2\sum_{i=1}^{n}\frac{1}{\sigma_{y_i}^2}(y_i - \widehat{a}x_i - \widehat{b}) & = & 0 \end{array}\right. \tag{7.20}$$

waaruit volgt

$$\left\{\begin{array}{rcl} \widehat{a}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_{y_i}^2} + \widehat{b}\sum_{i=1}^{n}\frac{x_i}{\sigma_{y_i}^2} & = & \sum_{i=1}^{n}\frac{x_i y_i}{\sigma_{y_i}^2} \\ \widehat{a}\sum_{i=1}^{n}\frac{x_i}{\sigma_{y_i}^2} + \widehat{b}\sum_{i=1}^{n}\frac{1}{\sigma_{y_i}^2} & = & \sum_{i=1}^{n}\frac{y_i}{\sigma_{y_i}^2} \end{array}\right. \tag{7.21}$$

We kunnen eenvoudig de uitdrukkingen voor $\widehat{a}$ en $\widehat{b}$ terugvinden door één vergelijking op te lossen naar $\widehat{b}$, de bekomen uitdrukking substitueren in de andere vergelijking, die we dan eenduidig kunnen oplossen naar $\widehat{a}$. Met behulp van de gevonden uitdrukking voor $\widehat{a}$ vinden we dan via de eerste vergelijking ook de uitdrukking voor $\widehat{b}$. In de cursussen Algebra zullen jullie een methode zien via de determinant van de coëfficiënten van de vergelijking. Beide methoden geven uiteraard hetzelfde resultaat. Hieronder vinden jullie de uitwerking voor de methode via de determinant. De determinant van het stelsel van de twee vergelijkingen 7.21 is gelijk aan

$$\Delta = \begin{vmatrix} \sum_{i=1}^{n} \frac{x_i^2}{\sigma_{y_i}^2} & \sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2} \\ \sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2} & \sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2} \end{vmatrix} = \left(\sum_{i=1}^{n} \frac{x_i^2}{\sigma_{y_i}^2}\right)\left(\sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2}\right)^2 . \tag{7.22}$$

waarmee we de waarde a en b kunnen berekenen voor schatters $\widehat{a}$ en $\widehat{b}$

$$\widehat{a} = \frac{1}{\Delta} \begin{vmatrix} \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_{y_i}^2} & \sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2} \\ \sum_{i=1}^{n} \frac{y_i}{\sigma_{y_i}^2} & \sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2} \end{vmatrix} = \frac{1}{\Delta}\left(\left(\sum_{i=1}^{n} \frac{x_i y_i}{\sigma_{y_i}^2}\right)\left(\sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2}\right)\left(\sum_{i=1}^{n} \frac{y_i}{\sigma_{y_i}^2}\right)\right) \tag{7.23}$$

en

$$\widehat{b} = \frac{1}{\Delta} \begin{vmatrix} \sum_{i=1}^{n} \frac{x_i^2}{\sigma_{y_i}^2} & \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_{y_i}^2} \\ \sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2} & \sum_{i=1}^{n} \frac{y_i}{\sigma_{y_i}^2} \end{vmatrix} = \frac{1}{\Delta}\left(\left(\sum_{i=1}^{n} \frac{x_i^2}{\sigma_{y_i}^2}\right)\left(\sum_{i=1}^{n} \frac{y_i}{\sigma_{y_i}^2}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2}\right)\left(\sum_{i=1}^{n} \frac{x_i y_i}{\sigma_{y_i}^2}\right)\right) \tag{7.24}$$

Met behulp van uitdrukking 6.35 kunnen we de onzekerheden op de schatters $\widehat{a}$ en $\widehat{b}$ bepalen. We starten met de uitdrukkingen

$$\begin{cases} \sigma_{\widehat{a}}^2 &= \sum_{j=1}^{n} \left(\frac{\partial \widehat{a}}{\partial y_j}\right)^2 \sigma_{y_j}^2 \\ \sigma_{\widehat{b}}^2 &= \sum_{j=1}^{n} \left(\frac{\partial \widehat{b}}{\partial y_j}\right)^2 \sigma_{y_j}^2 \end{cases} \tag{7.25}$$

waarin we

$$\begin{cases} \frac{\partial \widehat{a}}{\partial y_j} &= \frac{1}{\Delta}\left(\frac{x_j}{\sigma_{y_j}^2}\left(\sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2}\right) - \frac{1}{\sigma_{y_j}^2}\left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2}\right)\right) \\ \frac{\partial \widehat{b}}{\partial y_j} &= \frac{1}{\Delta}\left(\frac{1}{\sigma_{y_j}^2}\left(\sum_{i=1}^{n} \frac{x_i^2}{\sigma_{y_i}^2}\right) - \frac{x_j}{\sigma_{y_j}^2}\left(\sum_{i=1}^{n} \frac{x_i}{\sigma_{y_i}^2}\right)\right) \end{cases} \tag{7.26}$$

nodig hebben. Na het nodige rekenwerk bekomen we

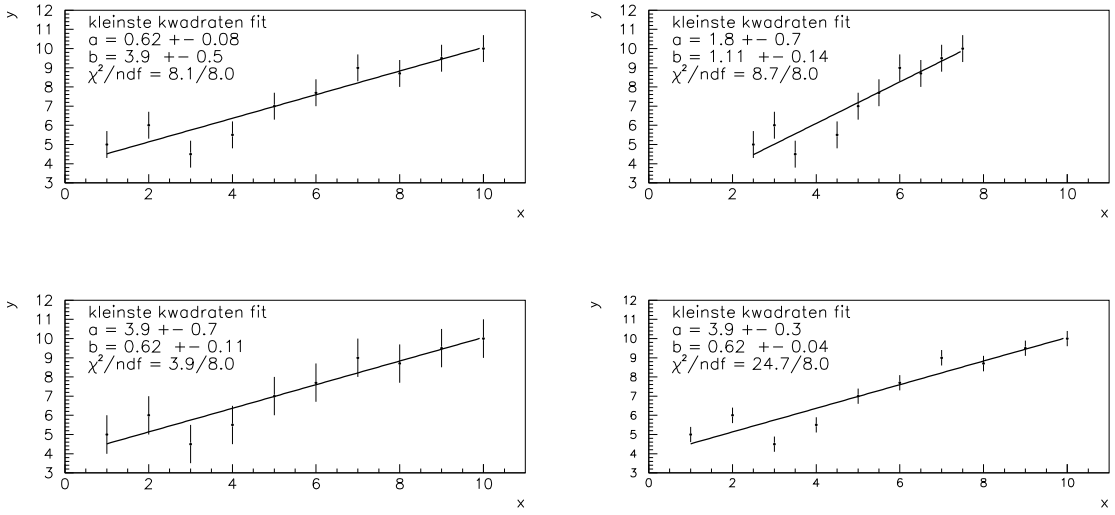
$$\begin{cases} \sigma_{\widehat{a}}^2 &= \frac{1}{\Delta} \sum_{i=1}^{n} \frac{1}{\sigma_{y_i}^2} \\ \sigma_{\widehat{b}}^2 &= \frac{1}{\Delta} \sum_{i=1}^{n} \frac{x_i^2}{\sigma_{y_i}^2} \end{cases} \tag{7.27}$$

als varianties op de coëfficiënten van de lineaire vergelijking. Voor sommige experimenten kan het gebeuren dat de onzekerheden σ_{y_i} gekend zijn voor alle n metingen. Het kan gebeuren dat deze niet gekend zijn en men ze bijgevolg moet schatten uit de n empirische

gegevens. Hiervoor moeten we wel veronderstellen dat voor alle n metingen, de onzekerheid gelijk is. Dit kunnen we doen aan de hand van de volgende schatter

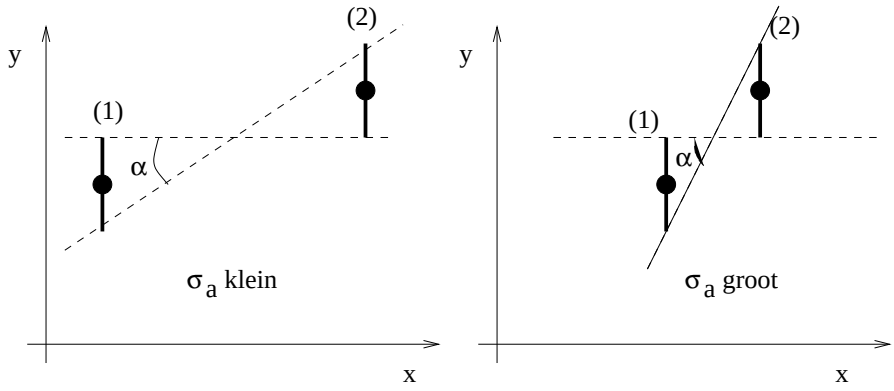
$$\widehat{\sigma_y}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b)^2 \tag{7.28}$$

waar de factor $(n-2)$ weergeeft dat we de gegevens gebruiken om twee parameters te schatten, namelijk $\widehat{a}$ en $\widehat{b}$.



Figuur 7.5: *Bepaling van de beste rechte, enkele voorbeelden.*

In Figuur 7.5 vinden we vier voorbeelden waar een beste rechte bepaald is uit een verzameling van 10 gegevens $\{(x_i, y_i, \sigma_{y_i}) \mid i \in \{1, 2, ..., 10\}\}$. Steeds worden de waarden van de schatters $\widehat{a}$ en $\widehat{b}$ weergegeven die de som van de kwadraten minimaliseren, alsook de waarde van Q^2 of χ^2 bij dit minimum. Voor de eenvoud hebben we de waarden σ_{y_i} voor alle $i \in \{1, 2, ..., 10\}$ gelijkgesteld. In de bovenste twee grafieken is $\sigma_{y_i} = 0.7$ in de onderste twee is $\sigma_{y_i} = 1.0$ (links) en $\sigma_{y_i} = 0.4$ (rechts). We zien duidelijk dat de waarde van σ_{y_i} geen invloed heeft op de beste waarden van $\widehat{a}$ en $\widehat{b}$. De minimale waarde van Q^2 of χ^2 daarentegen zal wel veranderen. Indien σ_{y_i} kleiner wordt, terwijl y_i gelijk blijft, zal de χ^2/ndf groter worden. We hebben hier de χ^2 gedeeld door het aantal vrijheidsgraden van de fit (ndf of '*number of degrees of freedom*') daar $\chi^2/ndf \simeq 1$ indien de fit goed is. Hier hebben we 10 metingen waaruit we twee parameters schatten, bijgevolg resten er ons $10 - 2 = 8$ vrijheidsgraden. Indien de theoretische afhankelijkheid tussen grootheid X en Y exact lineair is en de onzekerheden σ_{y_i} goed zijn ingeschat, zal de waarde van χ^2/ndf naar 1 convergeren indien n groot wordt. Indien de afhankelijkheid exact lineair is, maar de onzekerheden σ_{y_i} te groot of te klein zijn ingeschat, zal de verhouding χ^2/ndf afwijken van 1. Ze zal een waarde aannemen die respectievelijk kleiner of groter is dan 1. Dit wordt duidelijk door het bestuderen van de vergelijking van de kleinste kwadraten methode en bij het vergelijken van de verschillende grafieken in Figuur 7.5.



Figuur 7.6: *Illustratie voor de onzekerheid op de geschatte parameters.*

Daar we in een experiment meestal de waarden van x_i kunnen kiezen, moeten we de invloed bekijken van de verschillende mogelijkheden. De bovenste twee grafieken in Figuur 7.5 geven dezelfde waarden van y_i en σ_{y_i} weer, maar voor verschillende waarden van x_i. We zien dat de onzekerheid op de schatters $\widehat{a}$ en $\widehat{b}$ kleiner is, indien de meetpunten x_i verder uit elkaar liggen. Als men een experiment slechts een vast aantal keer kan herhalen, moeten we bijgevolg de ruimte van x_i zo groot mogelijk houden. Dit fenomeen wordt duidelijk door Figuur 7.6 te bekijken. Indien de twee punten x_1 en x_2 dichter tegen elkaar liggen, zal de mogelijk parameterruimte voor schatter $\widehat{a}$ veel groter worden [2], daar de openingshoek tussen de twee stippellijnen groter wordt. Er kunnen namelijk veel meer rechten met verschillende richtingscoëfficiënten doorheen de metingen lopen. Wordt die hoek α kleiner, dan is de parameterruimte voor schatter $\widehat{a}$ kleiner en kan men bijgevolg een betere schatting maken van de waarde van a_0.

[2] De schatter $\widehat{a}$ geeft de richtingsco¨effi ci¨ent van de beste rechte.

Hoofdstuk 8

Betrouwbaarheidsintervallen

"As far as the laws of mathematics refer to reality,
they are not certain; and as far as they are certain,
they do not refer to reality.'

Albert Einstein (1879-1955)

Indien de waarschijnlijkheidsdichtheidsverdeling van een schatter $\widehat{\theta}$ een normale of Gaussiaanse curve volgt, kunnen we een experimenteel resultaat noteren als de steekproefwaarde van de schatter en een schatting van de standaardafwijking (zie uitdrukking 6.29). Indien de verdeling van de schatter afwijkt van de normale of Gaussiaanse verdeling of indien er fysische grenzen zijn voor de waarde van de te schatten parameter, worden de resultaten van een experiment meestal voorgesteld als betrouwbaarheidsintervallen. Tot op vandaag is er geen perfecte methode om die betrouwbaarheidsintervallen op te stellen. Hieronder beschrijven we slechts één methode, namelijk die ontworpen door Jerzy Neyman (1894-1981) in 1937 (zie Figuur 8.1 voor een portret). Met behulp van de methode van Neyman bepalen we een interval dat met een zekere waarschijnlijkheid de theoretische waarde van een parameter bevat.

JERZY NEYMAN

Figuur 8.1: *Jerzy Neyman.*

8.1 Algemene definitie en interpretatie

Beschouw een stochastiek X die een theoretische verdeling $f_X(x \mid \theta)$ volgt, waar x het resultaat van een meting is en θ een parameter van de verdeling met onbekende waarde θ_0. We kunnen veronderstellen dat X een schatter is van parameter θ. Uitgaande van de kennis van $f_X(x \mid \theta)$ kunnen we voor een willekeurige waarschijnlijkheid α en een willekeurige waarde

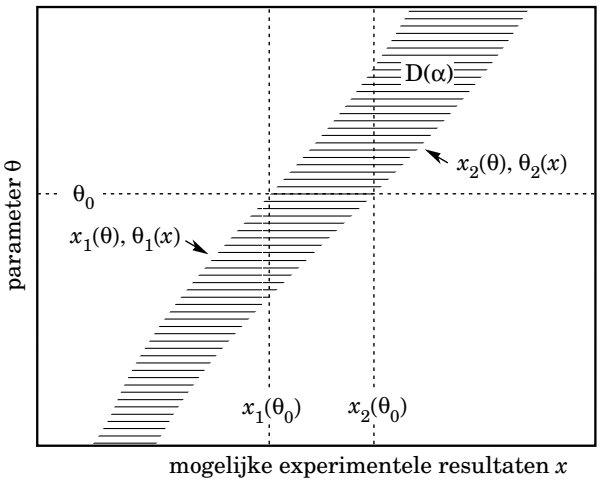
van de parameter θ, een interval $[x_1, x_2]$ vinden die $(1-\alpha)\%$ van de totale waarschijnlijkheid bevat. Dit kunnen we schrijven als

$$P(x_1 < x < x_2 \mid \theta) = 1 - \alpha = \int_{x_1}^{x_2} f_X(x \mid \theta) dx \tag{8.1}$$

met als algemene regel

$$P(x < x_1 \mid \theta) = P(x > x_2 \mid \theta) = \frac{\alpha}{2} = \int_{-\infty}^{x_1} f_X(x \mid \theta) dx = \int_{x_2}^{+\infty} f_X(x \mid \theta) dx \ . \tag{8.2}$$

Deze laatste regel of voorwaarde zorgt ervoor dat er evenveel kans is om een meting x te bekomen die links of rechts van het interval $[x_1, x_2]$ ligt. Figuur 8.2 illustreert het principe van de methode. Daar men de waarden x_1 en x_2 voor alle mogelijke waarden van θ en α kan bepalen, kunnen we spreken van de functies $x_1(\theta, \alpha)$ en $x_2(\theta, \alpha)$. In de figuur zien we de horizontale intervallen $[x_1(\theta, \alpha), x_2(\theta, \alpha)]$ voor verschillende waarden van θ. De unie van alle intervallen voor alle mogelijke waarden van θ noemen we de betrouwbaarheidsgordel $D(\alpha)$ en is enkel afhankelijk van de gekozen waarden van α. Met behulp van deze betrouwbaarheidsgordel wordt de parameterruimte voor θ gecorreleerd met de steekproefruimte waarin we X meten. Indien we een experiment uitvoeren met resultaat x_0 kunnen we in de grafiek een verticale lijn trekken bij $x = x_0$. Het betrouwbaarheidsinterval voor θ is de verzameling van alle waarden θ waarvoor het hiermee overeenstemmende segment $[x_1(\theta, \alpha), x_2(\theta, \alpha)]$ deze verticale lijn kruist. De waarden voor de parameter θ waar de verticale lijn de betrouwbaarheidsgordel verlaat noteren we met $\theta_1(x_0, \alpha)$ en $\theta_2(x_0, \alpha)$. De grootte van het betrouwbaarheidsinterval voor θ zal bijgevolg afhankelijk zijn van α. We spreken dan ook over een betrouwbaarheidsinterval met een betrouwbaarheidsniveau [1] gelijk aan $(1-\alpha)\%$. Hoe kleiner α, hoe groter het betrouwbaarheidsinterval voor θ.



Figuur 8.2: *Illustratie bij het begrip betrouwbaarheidsgordel.*

Stel nu dat de echte waarde van de parameter θ gelijk is aan θ_0. In de figuur kunnen we zien dat θ_0 in het betrouwbaarheidsinterval $[\theta_1(x, \alpha), \theta_2(x, \alpha)]$ ligt als en slechts als x in het interval $[x_1(\theta_0, \alpha), x_2(\theta_0, \alpha)]$ ligt. Beide gebeurtenissen hebben bijgevolg dezelfde waarschijnlijkheid en omdat dit geldig is voor alle waarden θ_0, bekomen we

$$1 - \alpha = P\left(x_1(\theta, \alpha) < x < x_2(\theta, \alpha)\right) = P\left(\theta_2(x, \alpha) < \theta < \theta_1(x, \alpha)\right) \tag{8.3}$$

[1] In het Engels worden die termen: confidence belt, confidence interval en confidence level soms afgekort met CL.

waar we overgaan van stochastiek X naar stochastieken $\theta_1(X)$ en $\theta_2(X)$. Indien we verschillende steekproeven nemen, zal het betrouwbaarheidsinterval $[\theta_1, \theta_2]$ veranderen en in $(1-\alpha)\%$ van de gevallen zal de theoretische constante waarde van θ erbinnen liggen. Hiermee hebben we het experimenteel resultaat geschreven als een conclusie over de theoretische waarde van de parameter zelf en met een duidelijke interpretatie. De betrouwbaarheidsgordel $D(\alpha)$ is een nuttige constructie om over te gaan van de parameter ruimte naar de meetruimte, of omgekeerd. Let erop dat we geen voorwaarden hebben op de vorm van de waarschijnlijkheidsdichtheidsverdeling $f_X(x \mid \theta)$. Wel moeten we ervan uitgaan dat deze functie gekend is.

In plaats van centrale intervallen te construeren door het opleggen van voorwaarde 8.2, is het soms nuttig om enkel het bovenste of onderste betrouwbaarheidsinterval te bepalen. We moeten de uitdrukking 8.2 vervangen door

$$\alpha = \int_{x_2}^{+\infty} f_X(x \mid \theta) dx \qquad \text{en} \qquad x_1 = -\infty \tag{8.4}$$

voor $(1-\alpha)\%$ bovenlimieten te bepalen van het interval, en

$$\alpha = \int_{-\infty}^{x_1} f_X(x \mid \theta) dx \qquad \text{en} \qquad x_2 = +\infty \tag{8.5}$$

voor $(1-\alpha)\%$ onderlimieten te bepalen van het interval.

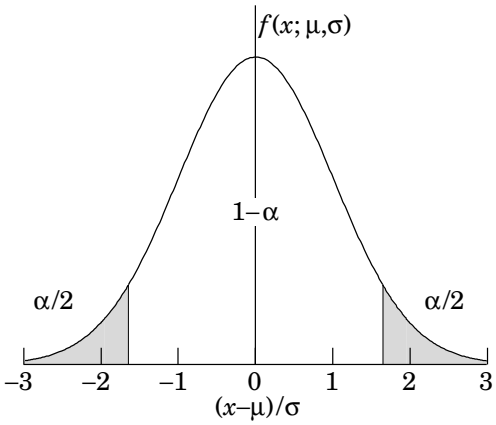
De definitie en interpretatie van de betrouwbaarheidsintervallen die we hier beschrijven, kunnen we eenvoudig uitbreiden naar meerdere dimensies.

8.2 Normaal verdeelde data

Voor de meeste metingen van theoretische parameters die men uitvoert, kan men een schatter opstellen die een normale of Gaussische verdeling volgt. Denk maar aan de centrale limietstelling. Bijgevolg is het nuttig om de betrouwbaarheidsintervallen voor deze categorie van schatters grondiger te bestuderen.

Stel dat we de theoretische verwachtingswaarde $E[\mu] = \mu_0$ van een normaal verdeelde schatter $\mu \sim N(\mu_0, \sigma_\mu^2)$ willen schatten [2]. Met behulp van vergelijking 8.1 bekomen we voor een willekeurige waarde van μ

$$1-\alpha = \frac{1}{\sigma_\mu \sqrt{2\pi}} \int_{\mu-\delta}^{\mu+\delta} e^{-\frac{1}{2}\left(\frac{\overline{x}-\mu}{\sigma_\mu}\right)^2} d\overline{x} \tag{8.6}$$



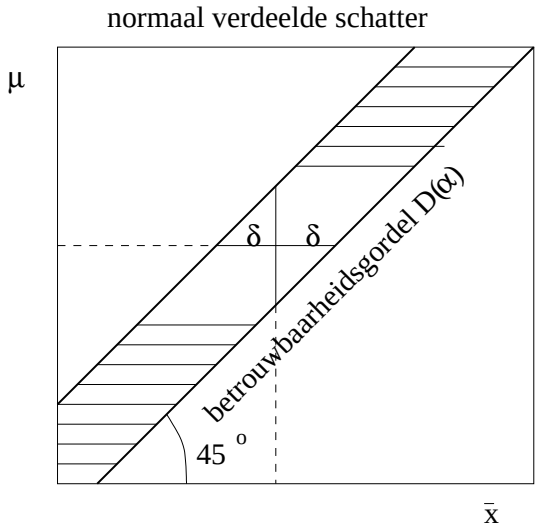
Figuur 8.3: *Betrouwbaarheidsintervallen met een normaal verdeelde schatter (90% CL).*

[2] De parameter θ uit vorige sectie wordt nu de verwachtingswaarde μ, terwijl de meting of schatter X uit vorige sectie nu het rekenkundig gemiddelde $\overline{x}$ wordt.

waar we veronderstellen dat de parameter σ_μ gekend is. Dit is de waarschijnlijkheid dat de gemeten waarde van $\overline{x}$ binnen het interval $[\mu - \delta, \mu + \delta]$ ligt. Voor een gegeven waarde van α kunnen we voor elke waarde van μ, de waarde van δ bepalen en hiermee de centrale betrouwbaarheidsgordel opstellen. Dit gaat uiteraard veel sneller met de nodige computerprogramma's. De normale verdeling is symmetrisch voor $\overline{x}$ en μ. Anders gezegd indien we beide grootheden $\overline{x}$ en μ omwisselen in de uitdrukking 8.6 zal de uitkomst $(1 - \alpha)$ niet veranderen. Bijgevolg mogen we ook zeggen dat de bekomen waarde van $(1 - \alpha)$ de waarschijnlijkheid weergeeft dat het interval $[\overline{x} - \delta, \overline{x} + \delta]$ de waarde μ bevat. Het toepassen van de methode van Neyman voor een normaal verdeelde schatter is dus eenvoudig, daar men de betrouwbaarheidsgordel niet moet construeren. De onzekerheid op de meting $\overline{x}$ kunnen we direct gebruiken als onzekerheid op de theoretische parameter μ

$$P(\mu - \delta < \overline{x} < \mu + \delta) = P(\overline{x} - \delta < \mu < \overline{x} + \delta) \ . \tag{8.7}$$

Figuur 8.3 geeft het betrouwbaarheidsinterval indien $\delta = 1.64\sigma_\mu$, wat gelijk staat aan $(1 - \alpha) = 90\%$. Indien we $\delta = \sigma_\mu$ kiezen, bekomen we een betrouwbaarheidsinterval voor de parameter μ met 1σ standaardafwijkingen, en $(1 - \alpha) = 68.27\%$. Een illustratie van een betrouwbaarheidsgordel voor een normaal verdeelde schatter vinden we in Figuur 8.4. Hier is de te schatten parameter $\theta = \mu$ gelijk aan de meting $\overline{x}$, en bijgevolg volgt de betrouwbaarheidsgordel de bisectrice tussen de parameter-as (μ) en de meting-as ($\overline{x}$). Dit komt eigenlijk door de symmetrie van de normale verdeling 8.6 tussen $\overline{x}$ en μ. Dit voorbeeld is bijgevolg een zeer specifiek geval, maar wel het meest voorkomende geval in de praktijk.



Figuur 8.4: *Betrouwbaarheidsgordel met een normaal verdeelde schatter.*

Door deze symmetrie kunnen we voor een normale verdeling enkele betrouwbaarheidsintervallen eenvoudig berekenen. In Tabel 8.8 vinden we voor een aantal waarden van α de bijhorende waarde voor δ, dit via vergelijking 8.6.

α (in %)		δ (inaantal σ)	α (in %)		δ (inaantal σ)
31.73	$\longrightarrow$	1	20	$\longrightarrow$	1.28
4.55	$\longrightarrow$	2	10	$\longrightarrow$	1.64
0.27	$\longrightarrow$	3	5	$\longrightarrow$	1.96
0.0063	$\longrightarrow$	4	1	$\longrightarrow$	2.58
0.000057	$\longrightarrow$	5	0.1	$\longrightarrow$	3.29
0.00000020	$\longrightarrow$	6	0.01	$\longrightarrow$	3.89

(8.8)

Deze lijst van getallen zal zeer nuttig blijken bij het opstellen van betrouwbaarheidsintervallen en kunnen jullie uiteraard zelf uitbreiden.

In de literatuur bestaan er verschillende alternatieve methoden om met behulp van de Neyman methode, betrouwbaarheidsintervallen te bepalen in moeilijke gevallen. Bijvoorbeeld voor schatters die geen normale verdeling volgen of in het geval dat we aan de grenzen van de parameterruimte zitten (cfr. het gewicht dat niet negatief kan zijn).

Slotwoord

Deze inleidende cursus over waarschijnlijkheid en statistiek maakt het mogelijk om de experimentele gegevens tijdens de labo's van de cursus 'Meten en experimenteren' correct te benaderen en nadien te interpreteren.

Bij het opstellen van deze syllabus heb ik mijn inspiratie gevonden in vele andere nota's en boeken, o.a. de cursusnota's van Prof.De Groen (Vrije Universiteit Brussel), Prof.de Wolf (Universiteit Antwerpen), Prof.Metzger (Universiteit Nijmegen), Prof.Schoukens (Vrije Universiteit Brussel), enzovoort. In de bibliografie vinden jullie ook enkele referentiewerken.

Het begrip van een likelihood die eventjes werd aangehaald in het eerste hoofdstuk van deze cursus, is een krachtig centraal begrip in de wetenschap van de statistische analyse van experimentele gegevens. In de volgende studiejaren is dit concept dan ook een essentieel element in de cursussen statistiek. De regel van Bayes gebruikt dergelijke likelihoodfuncties om informatie over de experimentele gegevens te transformeren in informatie over de theorie. Met het begrip van een likelihood kunnen we ook het concept 'informatie' inleiden, wat aanleiding zal geven tot de definitie van optimale informatie. Er bestaan verschillende technieken om zo optimaal mogelijk informatie te extrageren uit een empirische verzameling gegevens.

Ook hebben we doorheen de cursus verwezen naar de matrixnotatie. In het eerste Bachelor jaar zullen jullie leren rekenen met matrices en ook hun eigenschappen opstellen. Bijgevolg kan men in de hogere studiejaren bijvoorbeeld de kleinste kwadraten methode opschrijven in deze matrixnotatie. Zoals jullie zullen zien, heeft dit veel practische toepassingen.

In het tweede Bachelor jaar Fysica krijgen jullie de vervolgcursus 'Statistische verwerking van experimentele gegevens'.

Oefeningen

1. De leeftijd van de studenten in een klas van 25 is de volgende:

   19.0, 18.7, 19.3, 19.2, 18.9, 19.0, 20.2, 19.9, 18.6, 19.4, 19.3 18.8, 19.3, 19.2, 18.7, 18.5, 18.6, 19.7, 19.9, 20.0, 19.5, 19.4, 19.6, 20.0, 18.9

   Bereken met behulp van deze empirische gegevens het rekenkundig gemiddelde en de standaardafwijking. Doe dit ook indien je de leeftijd van de docent erbij telt, namelijk 37.0. Is het effect van deze extra meting groter op het kental van de locatie of op het kental van de spreiding? Bepaal ook de scheefheid en de kurtosis in beide gevallen.

2. Van 12 studenten worden de punten op het examen 'klassieke mechanica' en 'kwantum mechanica' vergeleken. We bekomen volgende vergelijking:

Klassiek	22	48	76	10	22	4	68	44	10	76	14	56
Kwantum	63	39	61	30	51	44	74	78	55	58	41	69

   Stel deze empirische gegevens voor in een twee-dimensionale scatterplot. Met behulp van deze grafiek kan je al een eerste ruwe schatting maken van de correlatie tussen beide resultaten, doe dit. Bereken dan exact het rekenkundig gemiddelde, de covariantie en de correlatie coëfficiënt.

3. Maak met behulp van Mathematica een histogram van 1000 random getallen die een uniforme verdeling hebben tussen 0 en 1. Ga na dat deze verdeling effectief uniform is. Maak nu een twee-dimensionaal histogram met behulp van opeenvolgende random getallen voor de coördinaten x en y. Is dit twee-dimensionaal histogram opnieuw uniform? Bepaal de correlatie coëfficiënt tussen x en y.

4. Toon aan dat men de scheefheid ook kan schrijven als:

$$\gamma_1(X) = \frac{1}{\sigma_x^3}\left(E[X^3] - 3E[X]E[X^2] + 2E[X]^3\right) \tag{8.9}$$

   Soms is het handiger om dit kental op deze manier te bepalen.

5. Beschouw een bundel van mesonen, die bestaat uit 90% pionen en 10% kaonen (beide zijn deeltjes die behoren tot de groep van mesonen). Deze bundel wordt gedetecteerd door een toestel dat pionen kan registreren maar geen kaonen. Op deze manier kunnen we de deeltjes in de bundel karakteriseren. In de praktijk lukt dit uiteraard niet perfect. Zo heeft de detector een efficiëntie van 95% om pionen te detecteren, en ook een waarschijnlijkheid van 6% om toevallig het signaal van een kaon te registreren. Als we nu een signaal registreren in onze detector, wat is de waarschijnlijkheid dat het een pion was? Als we geen signaal registreren, wat is de waarschijnlijkheid dat het een kaon was?

6. De Mongoolse moerasgriep is een zeldzame ziekte die artsen slechts in 1 uit 10000 patiënten verwachtten. De symptomen die steeds voorkomen zijn uitslag op de huid en een slaperig gevoel, soms (in 60% van de gevallen) hebben deze patiënten ook een extreme dorst, en soms (in 20% van de gevallen) beginnen ze extreem geweldadig te niezen. Uiteraard kunnen deze symptomen ook voorkomen bij personen die deze ziekte niet hebben. Zo heeft 3% van de patiënten huiduitslag, 10% is slaperig, 2% heeft extreme dorst en 5% heeft klachten over niezen. Deze waarschijnlijkheden kan men als onafhankelijk beschouwen.

   Toon aan dat indien je naar een arst gaat met al deze symptomen, dat de waarschijnlijkheid dat je de Mongoolse moerasgriep hebt 80% is. Wat is deze waarschijnlijkheid indien je al deze symptomen hebt, behalve het extreem niezen?

7. Stel dat een anti-raket systeem een efficiëntie van 99.5% heeft om inkomende raketten te stoppen. Wat is de waarschijnlijkheid dat dit systeem alle van 100 inkomende raketten stopt? Hoeveel raketten moet de aanvaller lanceren om een kans groter dan 0.5 te hebben dat ten minste één raket niet gestopt word door het anti-raket systeem? De aanvaller is agressief (wanneer niet), en wil bijgevolg dat, met een kans groter dan 0.5, ten minste twee raketten inslaan. Hoeveel raketten heeft de aanvaller hier minstens voor nodig?

8. Welke empirische meting van stochastische grootheid X geeft de kleinste onzekerheid, een verzameling van 10 metingen met een onzekerheid of resolutie van 1 mm of slechts één meting met een onzekerheid of resolutie van 0.2 mm ?

9. Een experiment waarvan de uitkomst een binomiaalverdeling volgt, geeft N_s geslaagde uitkomsten en N_f niet geslaagde uitkomsten. Toon aan dat

$$\widehat{p} = \frac{N_s}{N_s + N_f} \tag{8.10}$$

   een consistente, alsook een zuivere schatter is voor de individuele waarschijnlijkheid p van de binomiaal verdeling.

10. Een student is aan het liften langs een niet te drukke straat, gemiddeld passeert slechts 1 auto per minuut. De waarschijnlijkheid dat een autobestuurder de lifter meeneemt is 1%. Wat is de waarschijnlijkheid dat de student nog steeds aan het wachten is

    - nadat 60 auto's gepasseerd zijn?
    - na 1 uur wachten?

11. Beschouw een stochastiek die een normale verdeling volgt

    - Wat is de waarschijnlijkheid dat een waarde van de stochastiek meer dan 1.23σ afwijkt van de verwachtingswaarde?
    - Wat is de waarschijnlijkheid dat een waarde van de stochastiek meer dan 2.43σ hoger ligt dan de verwachtingswaarde?

- Wat is de waarschijnlijkheid dat een waarde van de stochastiek minder dan 1.09σ onder de verwachtingswaarde ligt?
- Wat is de waarschijnlijkheid dat een waarde van de stochastiek hoger dan 0.45σ onder de verwachtingswaarde ligt?
- Wat is de waarschijnlijkheid dat een waarde van de stochastiek meer dan 0.5σ maar minder dan 1.5σ afwijkt van de verwachtingswaarde?
- Wat is de waarschijnlijkheid dat een waarde van de stochastiek boven 1.2σ onder de verwachtingswaarde maar onder 2.1σ boven de verwachtingswaarde ligt?
- Binnen hoeveel standaardafwijkingen ligt 50% van alle mogelijke waarden van stochastiek X?
- Hoeveel standaardafwijkingen boven de verwachtingswaarde moeten we gaan om 99% van de waarden van stochastiek X eronder te hebben?

12. Gedurende een lawine van meteorieten, komen gemiddeld 15.7 meteorieten per uur aan. Wat is de waarschijnlijkheid om 5 meteorieten te observeren in een tijdspanne van 30 minuten? Welke waarde bekom je indien je de Poisson verdeling benadert met een normale verdeling?

13. Vier waarden worden geobserveerd uit een normale verdeling, namelijk 3.9, 4.5, 5.5 en 6.1. De verwachtingswaarde van de normale verdeling is gekend en gelijk aan 4.9. De variantie is echter onbekend.

    - Wat is de kans dat een volgende waarneming uit deze verdeling een waarde heeft groter dan 7.3?
    - Wat is de kans dat het gemiddelde van vier volgende waarnemingen tussen 3.8 en 6.0 ligt?

14. Toon aan dat voor n onafhankelijke stochastieken X die een uniforme verdeling volgen tussen 0 en 1, dat de waarschijnlijkheidsdichtheidsverdeling voor stochastiek G met

$$g = \frac{\sum_{i=1}^{n} x_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \tag{8.11}$$

    een normale verdeling benadert $G \sim N(0,1)$ indien $n \to \infty$. Toon dit ook aan via een simulatie in Mathematica, waar je verschillende waarden voor n kan gebruiken.

15. Beschouw 7 getallen die een normale verdeling volgen:

$$20.0,\ 19.7,\ 20.6,\ 18.5,\ 21.2,\ 20.8,\ 20.7 \tag{8.12}$$

    - Bereken het rekenkundig gemiddelde en de onzekerheid op dit gemiddelde indien de getallen een normale verdeling volgen met een standaardafwijking van 0.8.
    - Schat de standaardafwijking indien je weet dat de verwachtingswaarde van de normale verdeling gelijk is aan 20.0. Wat is de onzekerheid op deze schatting ?

- Schat de standaardafwijking zonder voorkennis over de verwachtingswaarde. Wat is nu de onzekerheid op deze schatting ?
- Bepaal de onzekerheid op het gemiddelde zonder voorkennis over de standaardafwijking van de normale verdeling.

16. Stel dat we het rekenkundig gemiddelde van x en y bepaald hebben met varianties σ_x^2 en σ_y^2. De covariantie tussen beide is gelijk aan nul. Bepaal nu de varianties en covarianties van r en θ indien

$$r^2 = x^2 + y^2 \qquad \text{en} \qquad tan\theta = \frac{y}{x} \tag{8.13}$$

Hier gaan we bijgevolg over van Cartesische coördinaten naar poolcoördinaten.

17. We meten $x = 10.0 \pm 0.5$ en $y = 2.0 \pm 0.5$. Wat is nu de onzekerheid op de waarde van x/y? Stel een simulatie op in Mathematica die de geldigheid van deze voortplanting van onzekerheden test.

18. Een object beweegt langs een traject met een constante maar ongekende snelheid. Het passeert opo een punt $d = 0$ op het exacte tijdstip $t = 0$. Op welbepaalde vaste tijdstippen, bepaald door een stroboscoop, kan je een foto nemen van het object en hiermee zijn positie bepalen met een onzekerheid van 2 mm. Je bekomt de volgende resultaten:

Tijd t (seconden)	1.0	2.0	3.0	4.0	5.0	6.0
Afstand d (mm)	11	19	33	40	49	61

Bepaal de snelheid v, alsook de onzekerheid op deze snelheid via de methode van de kleinste kwadraten. Maak ook een grafiek van de $\Delta\chi^2(v)$ en vergelijk de onzekerheid bekomen uit deze grafiek met diegene die je berekent hebt.

19. Een object beweegt langs een traject met een constante maar ongekende snelheid. Het passeert opo een punt $d = 0$ op het exacte tijdstip $t = 0$. Op welbepaalde vaste afstanden, bepaald door licht-sensoren langs het traject, kan je de tijd meten waarop het object deze vaste positie bereikt. Deze tijd kan je meten met een onzekerheid van 0.1 seconden. Je bekomt de volgende resultaten:

Tijd t (seconden)	1.1	2.2	2.9	4.1	5.0	5.8
Afstand d (mm)	10	20	30	40	50	60

Bepaal de snelheid v, alsook de onzekerheid op deze snelheid via de methode van de kleinste kwadraten. Maak ook een grafiek van de $\Delta\chi^2(v)$ en vergelijk de onzekerheid bekomen uit deze grafiek met diegene die je berekent hebt. Vergelijk de grafiek met diegene die je in vorige oefening bekomen hebt.

20. Je wil de versnelling g bepalen van een object te wijten aan de zwaartekracht. Het object laat je los met behulp van een elektrische magneet die je aan en af kan zetten. Via de uitdrukking $d = \frac{1}{2}gt^2$ willen we de versnelling g bepalen uit vast bepaalde afstanden d en een tijd t die je meet met een onzekerheid van 0.01 seconden. Je bekomt volgende resultaten:

Tijd t (seconden)	0.16	0.40	0.58	0.72	0.97
Afstand d (mm)	0.20	1.00	2.00	3.00	5.00

Bepaal de zogenaamde valversnelling g met bijhorende onzekerheid met behulp van bovenstaande gegevens. Beschouw ook de situatie waarin het magneetveld een constante maar ongekende tijd nodig heeft om uit te sterven en pas daarna de bal los te laten. Maak de grafiek van $\Delta\chi^2(g)$ en controleer je berekende onzekerheid.

Interpreteer het verschil tussen beide benaderingen.

21. Beschouw een radioactieve bron waarvan de straling gemeten word met een Geiger teller. De gegevens van deze teller worden elk uur geregistreerd voor een tijdsinterval van 1 minuut. Tijdens die ene minuut krijg je een aantal tellers. Je bekomt volgende resultaten na 0, 1, ..., 8 uur:

Tijd t (uur)	0	1	2	3	4	5	6	7	8
Aantal tellers	997	520	265	127	70	35	16	7	3

Gebruik deze gegevens om met behulp van de methode van de kleinste kwadraten de halfwaarde tijd $t_{\frac{1}{@}}$ van de radioactieve bron te bepalen. Maak de grafiek $\Delta\chi^2(t_{\frac{1}{2}})$ en controleer hiermee de onzekerheid op $t_{\frac{1}{2}}$.

22. Stel we hebben verschillende metingen gedaan van een stochastische grootheid Y, met resultaat 5 ± 2, 3 ± 1, 5 ± 1 en 8 ± 2 bij vier verschillende waarden van stochastische grootheid X, namelijk respectievelijk $x = -0.6, -0.2, 0.2, 0.6$. We veronderstellen een parabolisch verband tussen beide grootheden, namelijk

$$y(x) = \theta_1 + \theta_2 x + \theta_3 x^2 \tag{8.14}$$

Maak met behulp met de methode van de kleinste kwadraten een schatting voor de waarden van de parameters θ_i. Bepaal ook de onzekerheid op de geschatte waarde van de parameters. Bereken dan de waarde van y en zijn onzekerheid voor een waarde van $x = 1$.

Bibliografie

[1] M.G. Kendall and A. Stuart, *The advanced theory of statistics*, Griffin, Vol.I (1977)
dit werk bestaat uit in totaal drie volumes en wordt door de meesten onder ons aanzien als *het* naslagwerk in verband met statistiek

[2] W.T. Eadie *et al.*, *Statistical methods in experimental physics*, North-Holland (1971)
dit is een uitstekend boek voor fysici, maar wel van een hoger niveau

[3] R.J. Barlow, *Statistics: a guide to the use of statistical methods in the physical sciences*, Wiley (1989)
een gemakkelijk leesbaar boek en omvat de meeste begrippen die we hebben besproken

[4] het internet !!