Data Analyse en Statistiek

Hella Snoek & Marthe Schut

Inhoud

Module I

Deze week gaan we kijken naar het begrip lineaire regressie. Hoe kunnen we een functie fitten aan een set meetwaardes. Er komen verschillende concepten aan bod.

Allereerst gaan we kijken waarom we meeste verdelingen die we tegenkomen Normaal verdeeld zijn. Hiervoor behandelen we de centrale limietstelling. Daarna gaan we kijken naar de kleinste kwadraten methode en de χ^2 verdeling. Uiteindelijk introduceren we ook methodes om hypotheses te toetsen.

We werken in de werkcolleges aan de opdrachten van deze module M3. Je vindt in het schema wanneer je aan welke opdrachten werkt en wanneer je deze moet inleveren. Vergeet ook niet te kijken naar het oefenmateriaal voor de derde tussentoets. De derde tussentoets volgt aan het einde van het vierde hoorcollege.

Hoofdstuk 1

De Centrale Limietstelling

De Centrale Limietstelling (Engels: Central Limit Theorem of CLT) is zonder meer de meest belangrijke stelling in de statistiek en in data analyses.

De **Centrale Limietstelling** zegt dat als je n onafhankelijk stochasten x_j hebt, waarvan elke stochast zijn eigen verdeling heeft met gemiddelde μ_j en variantie σ_j^2 (die niet persé dezelfde hoeven te zijn!), de som van deze stochasten $\sum_j x_j$ een normale verdeling zal volgen met het gemiddelde $\sum_j \mu_j$ en de variantie $\sum_j \sigma_j^2$.

En als de som een normale verdeling volgt, dat geldt dat ook voor het gemiddelde!

Wat deze stelling eigenlijk zegt is dat als je een combinatie hebt van vele oorzaken van onzekerheden, de uiteindelijke verdeling de normale verdeling zal hebben. En het maakt dus niet uit hoe de onderliggende verdelingen van de onzekerheden die je combineert eruit zien. Behalve dat ze een gedefinieerd gemiddelde en variantie moeten hebben. De Centrale Limietstelling verklaart waarom zoveel grootheden Normaal zijn verdeeld. Meestal is er namelijk sprake van een combinatie van een grote hoeveelheid toevalligheden die een rol spelen bij de onzekerheid van een meting. Hetzelfde geldt vaak voor de eigenschappen van een populatie, de natuurlijke verdeling van deze eigenschappen zijn vaak ook Normaal verdeeld om dezelfde reden.

Denk maar eens aan de vorming van een zandkorrel of van een ster. Het is dan begrijpelijk dat de sterren in een bolhoop een Normale massa verdeling kennen. Of de grootte van de zandkorrel op een strand. Bij de vorming van een ster of zandkorrel zijn er ook vele toevalligheden die invloed hebben op de grootte van zo'n object.

Het bewijs van deze stelling is bijzonder ingewikkeld en zullen we hier niet behandelen. Eventueel kun je hierverder lezen over de bewijsstelling.

Als je goed leest staat er dat de stochasten zelf een verdeling kennen met een gemiddelde

μ en een variantie σ^2. Dat is een belangrijke voorwaarde. Wiskundig kun je laten zien dat bijvoorbeeld stochasten die volgens de Cauchy of Landau verdeeld zijn bij combinatie geen Normaal verdeling opleveren. Toch is die beperking niet heel groot. In de natuur zijn praktisch alle stochastische verdelingen beperkt en voldoen dus aan de Centrale Limietstelling.

Twee leuke video's die de Centrale Limietstelling illustreren vindt kun je hieren hiervinden.

De convergentie van de distributie naar de Normaal verdeling hangt af van de onderliggende stochastische verdelingen.

Hoofdstuk 2

Normaalverdeling

De Normaalverdeling is een van de drie belangrijkste distributies in de statistische data analyse. Samen met de Poisson verdeling en de χ^2 verdeling.
(De Poisson verdeling omdat het de onzekerheid op tel-experimenten beschrijft, de χ^2 wordt in de volgende hoofdstukken in deze module beschreven.) We hebben in het hoofdstuk De Centrale Limietstelling gezien waarom onzekerheden op waarnemingen zo vaak Normaal zijn verdeeld.

2.1 Poisson en Normaal

Voordat we verder gaan over de normaalverdeling bekijken we eerst kort de Poissonverdeling. We hebben in module 1 al even kort gezien hoe deze verdeeld is. De Poisson is uitermate belangrijk in experimenten omdat het de onzekerheid op tel experimenten beschrijft. Voor een verwachtingswaarde van λ vinden we een standaarddeviatie van $\sqrt{\lambda}$ en zoals we al eerder hebben gezien mogen we deze bij het uitvoeren van een experiment vaak zien als de onzekerheid op de verwachtingswaarde zelf.

$$P(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{2.1}$$

De Poisson verdeling is asymmetrisch, vooral voor lage waardes van λ. Voor grotere waardes van λ zien we dat de verdeling steeds symmetrischer is en ook steeds meer overeenkomsten vertoont met een normaal verdeling.

Om dit te visualiseren tonen we de twee functies over elkaar heen voor een waarde van $\lambda = 60$. Deze vergelijken we nu met de normaal verdeling met $\mu = 60$ en $\sigma = \sqrt{60}$.

Er blijven natuurlijk verschillen, zo is de Poissonverdeling een discrete verdeling, maar de grote gelijkenis verklaart wel waarom we, voor grotere waardes van λ gebruik mogen maken van vergelijkingen die eigenlijk alleen voor de Normale verdeling gelden. Zoals bijvoorbeeld de regels voor de foutenpropagatie.

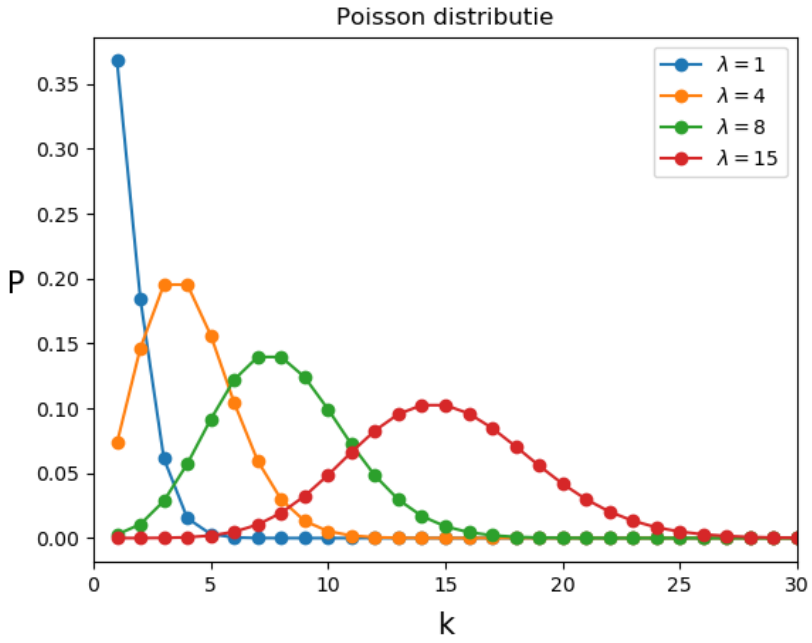

Figure 2.1:

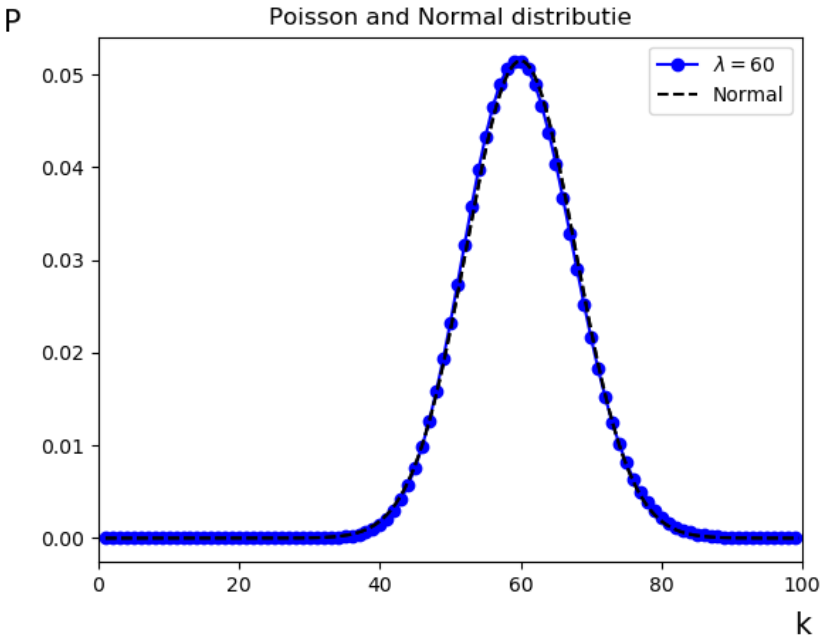

Figure 2.2:

2.2 Z-waardes en waarschijnlijkheden

We richten ons nu op de Normaalverdeling en herhalen nogmaals de vergelijking.

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$.

De functie heeft twee parameters, μ en σ, de notering is niet toevallig. De verwachtingswaarde van de normaal verdeling is precies μ en de standaarddeviatie is precies σ.

Hieronder zie je enkele voorbeelden van de Normale verdeling met verschillende waardes voor μ en σ.

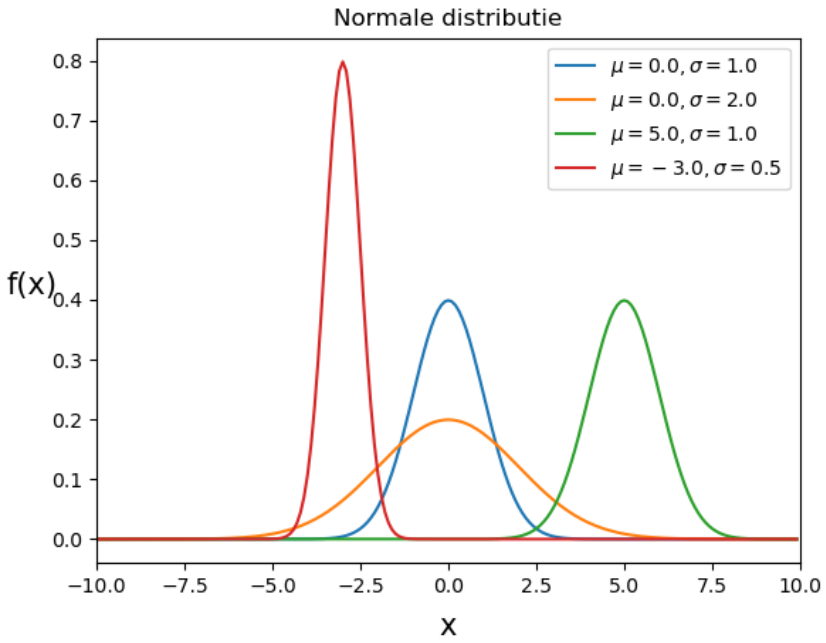

Figure 2.3:

We zien dat voor hogere waardes voor σ de datapunten meer verspreidt zullen zijn. Met andere woorden als de onzekerheid op een meting wordt uitgedrukt met de standaarddeviatie σ en de onzekerheid is groter, dat is de spreiding van de onderliggende kansdichtheidsverdeling ook groter.

Stel nu dat we een meting doen L en we kennen het populatiegemiddelde $\mu_L = 10.0$ cm met een spreiding van $\sigma_L = 2.0$ cm. De kans dat we een meting doen $L = 4.0$ cm is dan niet zo groot. Als de spreiding op het populatiegemiddelde daarentegen groter is, bijvoorbeeld $\sigma = 5.0$ cm dan is de kans veel groter om de meting van $L = 4.0$ cm te doen.

We kunnen dit uitdrukken met behulp van de Z-waarde ofwel Z-score.

Het oppervlak onder de normaalkromme behorende bij de kans om een waarde $X < x$ te vinden, is hieronder schematisch weergegeven:

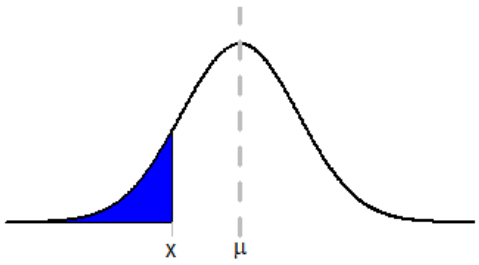

Figure 2.4:

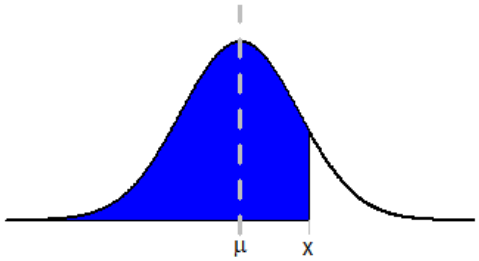

Figure 2.5:

Het oppervlak onder de normaalkromme behorende bij de kans om een waarde $X > x$ te vinden, is hieronder schematisch weergegeven:

Om dit oppervlak uit te rekenen gebruiken we de zogenoemde *Z-toets*. Stel een dataset met $n > 30$ datapunten is normaal verdeeld met gemiddelde μ en standaardafwijking σ. De $Z - score$, voor een bepaalde observatiewaarde x, is dan gelijk aan:

$$Z = \frac{x - \mu}{\sigma} \tag{2.2}$$

Stel een stochastische variabele X, met $n > 30$ datapunten, is normaal verdeeld met

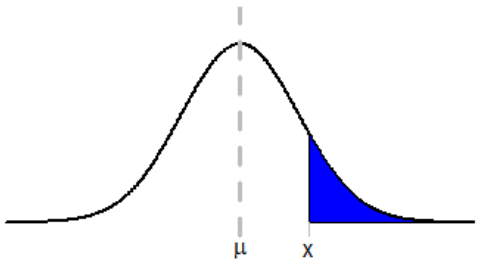

Figure 2.6:

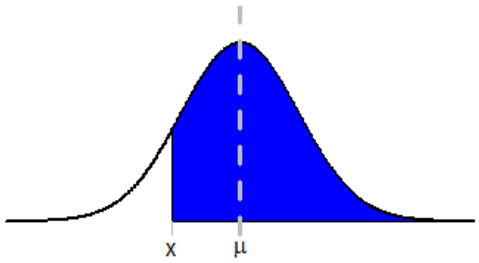

Figure 2.7:

gemiddelde μ en standaardafwijking σ.

Het oppervlak onder de normaalkromme, behorende bij de kans op een bepaalde waarde, hangt op de volgende manier van de z-score af.

De éénzijdige overschrijdingskans om een waarde $X < x$ te vinden is gelijk aan:

$$P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right) \tag{2.3}$$

De éénzijdige overschrijdingskans om een waarde $X > x$ te vinden is gelijk aan:

$$P(X > x) = 1 - P(X < x >) = 1 - P\left(Z < \frac{x-\mu}{\sigma}\right) \tag{2.4}$$

Dit kun je zelf nagaan door schetsen te maken van de bijbehorende oppervlakken onder de normaalkromme.

Bij de tweezijdige overschrijdingskans wordt de kans op een waarde groter dan de gestelde waarde opgeteld bij de kans op een waarde kleiner dan de gestelde waarde:

$$\begin{aligned} P(X = x) &= P(Z < \frac{x-\mu}{\sigma}) + P(Z > \frac{x-\mu}{\sigma}) \\ &= P(Z < \frac{x-\mu}{\sigma}) + \left(1 - P(Z < \frac{x-\mu}{\sigma})\right) \\ &= 2 \cdot P(Z < \frac{x-\mu}{\sigma}) - 1 \end{aligned} \tag{2.5}$$

Als je de z-score hebt berekend kun je uit de z-waarden tabel aflezen wat $P(Z < \frac{x-\mu}{\sigma})$ is.

Voorbeeld 1: Een stochast X is Normaal verdeeld met gemiddelde $\mu = 20$ en standaardafwijking $\sigma = 2$. Bereken de kans op een waarde $X < 16$.
Uitwerking: Het gaat hier om een eenzijdige overschijdingskans. Nu:

$$\begin{aligned} P(X < 16) &= P\left(Z < \frac{x-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{16-20}{2}\right) \\ &= P(Z < -2) \end{aligned}$$

Als we in de tabel kijken dan hoort er een waarde van 0.02275 bij deze Z-score.
Dus
$P(X < 16) = P\left(Z < \frac{x-\mu}{\sigma}\right) = 0.02275$
Er is in dit geval dus een kans van 2% dat we bij de gegeven dataset een waarde onder de 15 zullen vinden.
Voorbeeld 2: Een stochast X is normaal verdeeld met gemiddelde $\mu = 20$ en standaardafwijking $\sigma = 2$. Bereken de kans op een waarde $X > 22$.
Uitwerking: Het gaat hier om een eenzijdige overschijdingskans. Nu:

$$\begin{aligned} P(X > 22) &= 1 - P(X < 22 >) \\ &= 1 - P\left(Z < \frac{x-\mu}{\sigma}\right) \\ &= 1 - P\left(Z < \frac{22-20}{2}\right) \\ &= 1 - P(Z < 1) \end{aligned}$$

Als we in de tabel kijken dan hoort er een waarde van 0.84134 bij deze Z-score.
Dus
$P(X > 22) = 1 - P\left(Z < \frac{x-\mu}{\sigma}\right) = 1 - 0.84134 = 0.15866$
Er is in dit geval dus een kans van 15% dat we bij de gegeven dataset een waarde boven de 22 zullen vinden.

Hoofdstuk 3

De kleinste-kwadraten methode

De methode van de kleinste kwadraten is een manier om onbekende parameters te bepalen met behulp van een dataset. Het recept van de kleinste kwadraten is een heel krachtige schatter.

De methode van de kleinste kwadraten wordt ook wel lineaire regressie, of 'fitten' genoemd. De methode kan wiskundig worden afgeleid met behulp van de zogeheten 'maximale waarschijnlijkheid principes'.

3.1 Schatters

Eerst staan we nog even stil bij wat een *schatter* eigenlijk is. Vaak willen we met behulp van een meting een bepaalde grootheid te weten komen. Soms kunnen we die direct opmeten, maar vaak hebben we een methode of een recept nodig om dit te doen. Denk bijvoorbeeld bij de proef met de halfwaardedikte. We nemen eerst een set metingen en vervolgens hebben we een recept om hieruit de halfwaardedikte te bepalen. Deze halfwaardedikte *schatten* we met behulp van de methode die we een *schatter* noemen (Engels: estimator).

Als we een meting doen maken we altijd meetfouten, en als we een schatting doen dan is dus ook de nauwkeurigheid van de schatting begrensd. We hebben al een andere eigenschap van de schatter bekeken in de opgaves namelijk de onzuiverheid die wordt gegeven door:

$$b = E(\hat{x}) - \mu \tag{3.1}$$

Waarbij b de onzuiverheid is, $E(\hat{x})$ de verwachtingswaarde van de te schattengrootheid en μ het populatiegemiddelde van de te schatten grootheid.

3.2 De kleinste-kwadraten methode en χ^2

Een van de meest krachtige schatters is de methode van de kleinste-kwadraten. Deze gaan we hier bespreken.

Met de kleinste kwadraten methode minimaliseren we het kwadratisch verschil tussen een set metingen en de voorspelde waardes op die metingen, waarbij de voorspelling afhangt van één of meerdere parameters.

We beginnen meteen met een voorbeeld. Stel dat we een set metingen hebben die er als volgt uitziet.

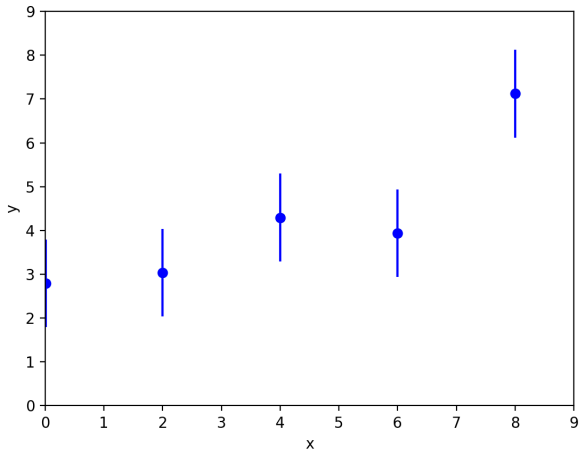

Figure 3.1:

We vermoeden een lineair verband tussen de variabelen x en y met parameters a en b. We willen nu een deze parameters schatten. De vraag is nu hoe bepalen we $\hat{a}$ en $\hat{b}$. Oftewel bij welke waardes van a en b wordt onze dataset optimaal beschreven met het lineaire verband.

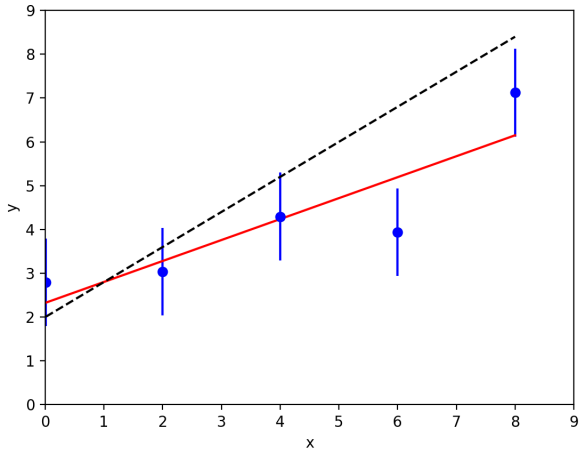

Figure 3.2:

Bekijk het voorbeeldje hierboven, we zien twee voorbeelden van oplossingen (de rode lijn

en de gestreepte zwarte lijn) met elk hun waardes voor a en b. De vraag is nu hoe bepaal je welke het beste is. Hiervoor gebruiken we een maat die we χ^2 noemen.

Stel dat we een functie $f(x; a, b)$ hebben die waardes van y voorspelt. En we hebben dataset met N waardes voor x : $x_1, x_2, ..., x_N$ met corresponderende waardes voor y : $y_1, y_2, ..., y_N$ waarbij elke waarde van y gemeten is met precisie σ_i. Nu kunnen we de som nemen van het kwadratische verschil van alle punten in de dataset met de voorspelde waardes $f(x_i; a, b)$, geschaald met de onzekerheden σ_i. Deze som noemen we χ^2 :

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i;a)}{\sigma_i} \right)^2$$

De meest optimale waarde voor a en b geeft de kleinste χ^2.

Door het kwadraat te gebruiken en niet het absolute verschil tussen de datapunten en de voorspelling geven we meer waarde aan de punten die ver van de voorspelling afliggen.

In de meeste gevallen kunnen we vaak algebraïsch de vergelijking oplossing. Namelijk door het minimum van de vergelijking te vinden. Als we nu kijken naar een functie die afhangt van slechts één parameter a dan kunnen we het minimum vinden op het punt dat de afgeleide gelijk is aan nul:

$$\frac{\partial \chi^2}{\partial a} = 0$$

Dit geeft:

$$\frac{\partial \chi^2}{\partial a} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f(x_i;a)}{\partial a} \left(y_i - f(x_i; a) \right) = 0$$

De betreffende waarde van a waarvoor dit geldt, genoteerd als $\hat{a}$ is dicht bij de echte waarde van de parameter, maar zal natuurlijk nog steeds een afschatting zijn.

In de vergelijkingen hierboven hebben we maar één afhankelijke parameter gezien maar dit principe kun je ook toepassen op functies met meerder afhankelijke parameters die je dan tegelijkertijd oplost.

Met behulp van een computerprogramma kun je het minimum ook vinden door de χ^2 voor veel waardes van a en b uit te rekenen en uit deze set van waardes het punt met de laagste χ^2 te bepalen. Uiteraard werkt dat ook voor functies die nog meer vrije parameters kennen.

Twee filmpjes die het principe van de kleinste kwadraten goed illustreren vind je hieren hier.

Om in te schatten **hoe goed** je fit gelukt is moeten we eerst meer weten over de χ^2-distributie. Daar gaat het volgende hoofdstuk over.

In opgave M3.1 ga je het principe van de kleinste kwadraten toepassen.

3.3 De χ^2 distributie

De χ^2 distributie is een maat voor het verschil tussen de voorspelde waardes en het kwadraat van het verschil. Als de functie f de data goed beschrijft zal de χ^2 klein zijn. Als de χ^2 dus groot blijft na het optimaliseren van de parameters van f dan is er duidelijk iets misgegaan. Het kan zijn dat de functie f de datapunten niet goed *kan* beschrijven, maar het kan ook zijn dat als je minimalisatie uitvoert met een computer deze het minimum niet goed heeft weten te vinden. Als daarentegen de χ^2 heel klein is gaat er ook iets mis. Waarschijnlijk heb je de onzekerheden op de datapunten heel erg overschat. Hoe groot je verwacht dat de waarde van χ^2 is na het optimaliseren van de kleinste kwadraten methode kun je bepalen. We gaan hieronder daar verder op in.

We hebben gezien in het hoofdstuk over de kleinste kwadraten methode, dat de χ^2 gedefinieerd is als het kwadratische gewogen verschil tussen de meetwaardes en de voorspelde waardes:

$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i;\hat{a},\hat{b},..)}{\sigma_i} \right)^2$.

Let op dat we hier de geoptimaliseerde parameters $\hat{a}$ van de functie hebben ingevuld. Deze waarde voor χ^2 is dus al geminimaliseerd voor de paramaters van f.

De χ^2 verdeling is een kansdichtheidsverdeling, en voldoet dus ook aan de voorwaardes hiervan. De functie ziet er als volgt uit:

$$P(\chi^2; df) = \frac{2^{-df/2}}{\Gamma(df/2)} \chi^{n-2} e^{-\chi^2/2}. \tag{3.2}$$

De Γ in de noemer is een speciale wiskundige functie. Deze zal pas in jullie tweede jaar volledig worden uitgelegd. Op dit moment kun je hem simpelweg interpreteren als een functie waar een normalisatie term uitkomt. Het is best een gekke functie, voorbeelden van uitkomsten: $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$ en $\Gamma(3/2) = 1/2\sqrt{\pi}$. Als je al meer wilt weten over de Γ-functie dan kun je daar bijvoorbeeld hier meer over lezen.

Zoals je ziet hangt de χ^2 kans ook af van een parameter df, dit is het aantal meetpunten, n, gereduceerd met het aantal parameters van de functie f. We noemen df het aantal *vrijheidsgraden*.

Voorbeeld Stel we hebben 10 meetwaardes en we gebruiken de kleinste kwadraten methode om 2 parameters van een functie f te optimaliseren. We hebben dan $df = 10 - 2 = 8$ vrijheidsgraden.

Hieronder zie je hoe de χ^2 eruit ziet voor verschillende waardes van df.

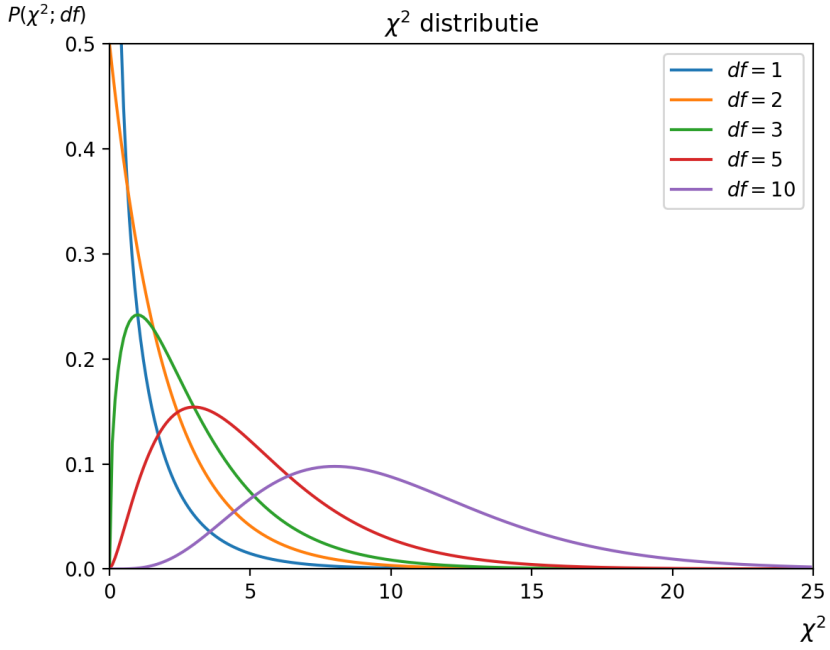

Figure 3.3:

De χ^2 distributie heeft een gemiddelde $\mu = df$ en een variantie van $var = 2df$. We verwachten dus een χ^2 van ongeveer **1 per vrijheidsgraad** te vinden. Als de χ^2 per vrijheidsgraad veel afwijkt van 1 dan is het waarschijnlijk dat er een probleem is met de fit. Het kan zijn dat de functie de relatie tussen de datapunten niet goed beschrijft, of dat er iets mis is met de onzekerheden op de datapunten.

3.4 Akaike Informatie Criterium

Stel dat je een dataset hebt waarvan je niet zeker weet door welke functie deze wordt beschreven. Je probeert twee functies uit, f_1 en f_2. En je minimaliseert voor beide functies de χ^2, deze zijn dan χ^2_1 en χ^2_2. Als algemene vuistvuistregelregel geldt dat de functie met de kleinste geminimaliseerde χ^2/df het beste de data beschrijft. Als in dat geval de betreffende χ^2/df dicht bij 1 ligt werkt deze vuistregel goed.

Voorbeeld 1 Stel dat we een dataset hebben met 10 gemeten waardes. We proberen twee functies uit:
$f_1(x; a, b) = a \cdot x + b$ en $f_2(x; a) = a \cdot x$
Als geminimaliseerde χ^2 voor de twee functies vinden we: $\chi_1^2 = 4.0$ en $\chi_2^2 = 13.0$.
De χ^2 per vrijheidsgraad is voor de twee functies:
$\chi_1^2/\text{vrijheidsgraad} = 4.0/(10-2) = 0.5$ en
$\chi_2^2/\text{vrijheidsgraad} = 13.0/(10-1) = 1.44$.
Op basis van de vuistregel zou je functie f_1 kiezen.
Voorbeeld 2 Stel dat we een dataset hebben met 10 gemeten waardes. We proberen twee functies uit:
$f_1(x; a, b) = a \cdot x + b$ en $f_2(x; a) = a \cdot x$
Als geminimaliseerde χ^2 voor de twee functies vinden we: $\chi_1^2 = 6.0$ en $\chi_2^2 = 9.0$.
De χ^2 per vrijheidsgraad is voor de twee functies:
$\chi_1^2/\text{vrijheidsgraad} = 6.0/(10-2) = 0.75$ en
$\chi_2^2/\text{vrijheidsgraad} = 9.0/(10-1) = 1.0$.
Op basis van de vuistregel zou je functie f_1 kiezen.

Als deze echter veel kleiner is dan 1 dan kun je betwijfelen of de bijbehorende functie wel echt de beste is. Beter is om dan het Akaike Informatie Criterium kun je gebruiken om uit te vinden welke functie het beste aan een dataset fit. Stel dat je een dataset hebt waarbij je n meetwaardes hebt die je beschreven hebt met een functie met p vrije parameters met een geminimaliseerde χ^2. Dan heeft het Akaike Informatie Criterium de volgende waarde:

$$AIC = \chi^2 + 2p + \frac{2p(p+1)}{n-p-1}. \tag{3.3}$$

Als we deze AIC berekenen voor beide functies dan is de functie met de laagste AIC de meest optimale.

Voorbeeld 1 Stel dat we een dataset hebben met 10 gemeten waardes. We proberen twee functies uit:
$f_1(x; a, b) = a \cdot x + b$ en $f_2(x; a) = a \cdot x$
Als geminimaliseerde χ^2 voor de twee functies vinden we: $\chi_1^2 = 4.0$ en $\chi_2^2 = 13.0$.
De AIC waarde voor de twee functies zijn nu:
$AIC_1 = 4.0 + 4 + 12/7 = 9.7$
$AIC_2 = 13.0 + 2 + 4/8 = 15.5$
Op basis van het Akaike Informatie criterium zou je functie f_1 kiezen.
Voorbeeld 2 Stel dat we een dataset hebben met 10 gemeten waardes. We proberen twee functies uit:
$f_1(x; a, b) = a \cdot x + b$ en $f_2(x; a) = a \cdot x$
Als geminimaliseerde χ^2 voor de twee functies vinden we: $\chi_1^2 = 6.0$ en $\chi_2^2 = 9.0$.
De χ^2 per vrijheidsgraad is voor de twee functies:
$AIC_1 = 6.0 + 4 + 12/7 = 11.8$
$AIC_2 = 9.0 + 2 + 4/8 = 11.5$
Op basis van de vuistregel zou je functie f_2 kiezen.

Hoofdstuk 4

Hypothese toetsen

Als je een steekproef hebt genomen en je wilt hiermee iets kunnen zeggen over de populatie dan moet er ook nagegaan worden in hoeverre de steekproef ons idee over de populatie ondersteund.

Dit wordt hypothese toetsen genoemd. Bij hypothese toetsen doorloop je de volgende stappen:

1. Hypothese opstellen
2. Significantielevel kiezen
    - Let op! Dit is iets anders dan de significantie waarin je een meetwaarde noteert.
3. p-waarde bepalen
4. Conclusie trekken

Deze stappen worden hieronder toegelicht.

4.1 Hypothese opstellen

Een hypothese is een uitspraak over een bepaalde eigenschap van een populatie. Je weet nog niet of deze uitspraak correct is. Een hypothese wordt geformuleerd als stelling.

Voorbeelden van hypotheses:

- 20% van de auto's in Nederland is blauw.
- 50% van de Nederlanders heeft blauwe ogen
- De valversnelling heeft als waarde in Nederland 9.81 ms^{-2}
- Studenten in Amsterdam halen hogere cijfers dan studenten in Groningen

Bij hypothese toetsen is er sprake van twee hypotheses. De zogenoemde *nulhypothese* en de *alternatieve hypothese*.

Bij hypothese toetsen wordt eerst aangenomen dat de eigenschap die onderzocht wordt niet waar is. Dit wordt de *nulhypothese* genoemd. De stelling dat de gewenste eigenschap wel waar is wordt de *alternatieve hypothese* genoemd. De nulhypotese wordt aangegeven met H_0, de alternatieve hypothese met H_α (ook H_1 is veelvoorkomend).

De procedure bij hypothese toetsen is dat je in eerste instantie aanneemt dat de eigenschap niet waar is (dus je houdt de nulhypothese aan) en dan onderzoekt of dit standhoudt in het kader van de gevonden resultaten. Uiteindelijk hoop je dat je de nulhypothese kunt verwerpen waardoor de alternatieve hypothese (en dus de gewenste waarde van de eigenschap) kunt aannemen.

Dus:

- Alternatieve hypothese H_α: De hypothese die zegt wat je verwacht te vinden in de dataset
- Nulhypothese: Het omgekeerde van de alternatieve hypothese.

Onderstaand de eerdere hypothesen met bijbehorende nulhypothesen:

Voorbeelden van alternatieve hypothese en nulhypothese:

- H_α: 20% van de auto's in Nederland is blauw.
- H_0: Het percentage blauwe auto's in Nederland is geen 20%.

- H_α: Meer dan 20% van de auto's in Nederland is blauw.
- H_0: Minder dan 20% van de auto's in Nederland is blauw.

- H_α: Het percentage Nederlanders met blauwe ogen is 50%.
- H_0: Het percentage Nederlanders met blauwe ogen is geen 50%.

- H_α: De valversnelling heeft als waarde in Nederland 9.81 ms^{-2}.
- H_0: De valversnelling in Nederland is niet gelijk aan 9.81 ms^{-2}.

- H_α: Studenten in Amsterdam halen hogere cijfers dan studenten in Groningen
- H_0: De studenten in Amsterdam halen lagere cijfers dan de studenten in Groningen.

- H_α: Het aantal katten in Nederland is groter dan 20 000
- H_0: Het aantal katten in Nederland is kleiner of gelijk aan 20 000

- H_α: Het percentage mensen over de gehele wereld met een hond is kleiner dan 40%
- H_0: Het percentage mensen over de gehele wereld met een hond is groter of gelijk aan 40%

In alle bovenstaande gevallen is het dus de procedure om te kijken of we genoeg bewijs hebben om de nulhypothese te kunnen verwerpen zodat we de alternatieve hypothese kunnen aannemen.

Extra opmerking: De term nulhypothese komt van het Engels 'null hypothesis' en de naamgeving slaat op de hypothese die verworpen (oftewel 'nullified') moet worden.

4.2 Significantielevel kiezen

De volgende stap in hypothese toetsen is het kiezen van het significantielevel. Dit houdt in dat we bepalen hoe zeker we ervan willen zijn dat we de correcte conclusie trekken, zonder precisie te verliezen.

Niet elke steekproef zal daadwerkelijk iets kunnen zeggen over de bijbehorende populatie. Als we bijvoorbeeld willen weten of het klopt dat 20% van de auto's in Nederland de kleur blauw heeft, maar in de steekproef kiezen we toevallig alleen auto's met een andere kleur, dan zouden we als conclusie kunnen trekken dat er in Nederland geen blauwe auto's rondrijden. Dit klopt echter niet met de daadwerkelijke populatie. Als je op de weg rijdt zie je namelijk wel degelijk blauwe auto's voorbij komen.

In het bovenstaande geval is de alternatieve hypothese dat 20% van de Auto's in Nederland blauw is maar we trekken de conclusie dat de nulhypothese (het percentage blauwe auto's in Nederland is geen 20%) correct is.

Er bestaat dus de kans dat we de berekeningen en statistiek op de juiste manier uitvoeren, maar alsnog de verkeerde conclusie trekken doordat de steekproef niet representatief is.

Er zijn twee manieren waarop de juiste conclusie wordt getrokken:

- De nulhypothese is correct en we concluderen ook daadwerkelijk vanuit de data dat deze correct is.
- De nulhypothese is niet correct en we concluderen ook daadwerkelijk vanuit de data dat we deze mogen verwerpen.

Omdat we de eigenschap alleen van de steekproef bekijken en niet van de gehele populatie weten we nooit helemaal zeker of we wel de juiste conclusie hebben getrokken (je weet immers niet of de nulhypothese in het echt correct/incorrect is).

Het zogenoemde *significantielevel* α geeft aan welk risico we willen lopen dat we de nulhypothese foutief verwerpen (d.w.z. de nulhypothese is eigenlijk wel waar maar we concluderen vanuit de data dat deze niet waar is).

Doorgaans wordt er voor het significantielevel gekozen uit de volgende drie waarden:

- $\alpha = 10\%$
- $\alpha = 5\%$
- $\alpha = 1\%$

Als de waargenomen kans (zie p-waarde hierna) kleiner is of gelijk aan het gekozen significantielevel a dan verwerpen we de nulhypothese. Is de waargenomen kans groter dan α dan verwerpen we de nulhypothese niet.

Kiezen we bijvoorbeeld een significantielevel van $\alpha = 5\%$ dan verwerpen we de nulhypothese zodra de waargenomen kans kleiner is dan 5%. Is de waargenomen kans groter dan 5%, dan verwerpen we de nulhypothese niet.

Hoe kleiner de kans is op de nulhypothese des te zekerder we ervan kunnen zijn dat we deze rechtmatig verwerpen. In principe wil je het significantieleven daarom zo laag mogelijk kiezen. Maar het kiezen van $\alpha = 1\%$ heeft een nadeel. Hoe kleiner het significantielevel des te groter de meetonzekerheid op de gemeten eigenschap. Het is dus altijd een afweging tussen het zo zeker mogelijk zijn van correctheid van het verwerpen van de nulhypothese, en het zo klein mogelijk houden van de meetonzekerheid op de waarde van de eigenschap.

4.3 p-Waarde bepalen

Na het kiezen van het significantielevel, bepalen (of meten) we de *p-waarde* behorende bij de nulhypothese. De p-waarde is de kans om de geobserveerde meetwaarden te vinden onder de aanname dat de nulhypothese correct is.

Stel we hebben de nulhypothese dat het percentage blauwe auto's in Nederland geen 20% is. We doen een meting waarbij we gedurende een dag het aantal blauwe auto's tellen die op de A6 voorbij komen. De kans dat we een uitkomst kunnen hebben van 25% blauwe auto's, **onder de aanname dat de nulhypothese correct is** (geen 20% blauwe auto's), is de p-waarde. Hoe kleiner de p-waarde die we vinden des te meer grond we hebben om de nulhypothese te verwerpen.

Er zijn verscheidene methodes voor het hypothese toetsen. In deze sectie behandelen we het bepalen van de p-waarde voor een normaal verdeelde dataset, middels de zogenoemde *z-toets*.

Ook voor data met een andere distributie kan de p-waarde bepaald worden via de z-toets voor een normale verdeling. Wel moet er dan een voldoende aantal metingen gedaan zijn zodat de **wet van grote aantallen** toegepast kan worden, en de data benaderd kan worden met een normale verdeling.

Afhankelijk van de manier waarop de nulhypothese en alternatieve hypothese opgesteld zijn, bepalen we de *eenzijdige overschrijdingskans* of de *tweezijdige overschrijdingskans*. Is de nulhypothese opgesteld met de formulering 'is gelijk aan' of 'is ongelijk aan', dan bepalen we de tweezijdige overschrijdingskans. Is de nulhypothese opgesteld met de formulering 'groter/kleiner dan' of 'groter/kleiner of gelijk aan' dan is het noodzakelijk om de eenzijdige overschrijdingskans te bepalen. Dus:

H_0 met	H_α met	type overschrijding
$=$	$\neq$	tweezijdig
$\neq$	$=$	tweezijdig
$\leq$	$>$	eenzijdig
$\geq$	$<$	eenzijdig

Voorbeelden van nulhypothesen waarbij er sprake is van het bepalen van de tweezijdige overschrijdingskans:

- H_0: Het percentage blauwe auto's in Nederland is geen 20%.
- H_0: Het percentage Nederlanders met blauwe ogen is 50%.

Voorbeelden van nulhypothesen waarbij er sprake is van het bepalen van de eenzijdige overschrijdingskans:

- H_0: De studenten in Amsterdam halen lagere cijfers dan de studenten in Groningen.
- H_0: Het aantal katten in Nederland is kleiner of gelijk aan 20 000
- H_0: Het percentage mensen over de gehele wereld met een hond, is groter of gelijk aan 40%

Zoals eerder vermeld geeft de p-waarde de kans dat waargenomen uitkomst gevonden kan worden onder de aanname dat de nulhypothese correct is. De p-waarde is dus gelijk aan een zeker oppervlak onder de normaalkromme. Deze kun je berekenen met de z-waarde.

4.4 Conclusie trekken

Tot nu toe hebben we de nulhypothese en de alternatieve hypothese opgesteld. Daarna hebben we bepaald welk significantielevel we zullen aanhouden. Vervolgens hebben we de z-score en daarmee de p-waarde bepaald. Maar hoe trek je aan de hand hiervan nu een conclusie over de nulhypothese?

Dit bekijken we aan de hand van een paar voorbeelden:

Voorbeeld 1: We onderzoeken de staartlengte van volgroeide lapjeskatten in Nederland, en stellen de volgende hypotheses op:

- H_α: De lengte van de staart van een volgroeide lapjeskat in Nederland is groter dan 10 cm.
- H_0: De lengte van de staart van een volgroeide lapjeskat in Nederland is kleiner of gelijk aan 10 cm.

Bij voorbaat kiezen we als significantielevel $\alpha = 5\%$.
We meten de staartlengte van 300 lapjeskatten in Nederland (met alle gevolgen van dien voor de onderzoekers), en zetten het resultaat uit in een histogram. Dit resulteert in een normale verdeling met gemiddelde $\mu = 25$ cm en een standaardafwijking 5 cm.
De nulhypothese stelde dat de lengte van de staart van een volgroeide lapjeskat kleiner of gelijk is aan 10 cm. We bepalen dus de p-waarde die hierbij hoort:

$$\begin{aligned} P(X < 10) &= P(Z < \frac{10 - 25}{5}) \\ &= P(Z < -3) \end{aligned}$$

Als we in de tabel kijken dan hoort er een waarde van 0.00135 bij deze Z-score. Dus: $P(X < 10) = P(Z < -3) = 0.00135$
De p-waarde is dus 0.14%. Op grond van het eerder gekozen significantielevel van 5% verwerpen we de nulhypothese. In dit geval is het zo dat we de nulhypothese ook hadden verworpen als we $\alpha = 10\%$ of $\alpha = 1\%$ hadden gekozen.

Is de p-waarde kleiner dan het gekozen significantielevel dan verwerpen we de nulhypothese. Is de p-waarde groter dan het gekozen significantielevel dan verwerpen we de nulhypothese niet.

Het is goed om te beseffen dat we **niet** kunnen zeggen dat onze alternatieve hypothese correct is of dat de nulhypothese fout is. De p-waarde geeft namelijk geen bewijs. Wel hebben we met de p-waarde een onderbouwing om de nulhypothese, met inachtname van het gekozen significantielevel, wel/niet te verwerpen.

Voorbeeld 2: We onderzoeken de gemiddelde lengte van alle vrouwen (> 18 jaar) in Nederland, en stellen de volgende hypotheses op:

- H_α: De gemiddelde lengte van alle vrouwen boven de 18 jaar is hoger dan 180 cm.
- H_0: De gemiddelde lengte van alle vrouwen boven de 18 jaar is lager dan of gelijk aan 180 cm.

Bij voorbaat kiezen we als significantielevel $\alpha = 5\%$.
We meten de lengte van 500 Nederlandse vrouwen boven de 18 jaar. De resultaten volgen een normale verdeling met gemiddelde $\mu = 165$ cm en een standaardafwijking 10 cm.
De nulhypothese stelde dat de gemiddelde lengte van de Nederlandse vrouwen hoger is dan 180 cm. We bepalen dus de p-waarde die hierbij hoort:

$$\begin{aligned} P(X > 180) &= 1 - P(X < 180) \\ &= 1 - P(Z < \frac{180 - 165}{10}) \\ &= 1 - P(Z < 1.5) \end{aligned}$$

Als we in de tabel kijken dan hoort er een waarde van 0.93319 bij deze Z-score. Dus: $P(X > 180) = 1 - P(Z < 1.5) = 0.06681$
De p-waarde is dus 6.7%. Op grond van het $\alpha = 5\%$ significantielevel verwerpen we de nulhypothese dus niet.

Hoofdstuk 5

Opdrachten module 3

Tijdens laptopcolleges 5 en 6 werken we aan het de opdrachten in module 3. In deze module gaan we werken aan twee opdrachten.

- M3.1 Grote Aantallen III ****
- M3.2 Halfwaardedikte III ***

De sterren geven een indicatie voor hoeveel werk een opdracht is.

De antwoorden van de opdrachten moet je invoeren in dit template en als **pdf** bestand inleveren via ANS. De deadlines voor de inleveropdrachten en informatie over ANS kun je hier vinden.

Als je vragen hebt, stel deze dan aan de assistent of stuur een email naar de coördinator. Veel succes!

5.1 M3.1 Grote Aantallen III ****

In deze opdracht gaan we het eindresultaat van M2.1 'fitten' met de kleinste kwadraten methode.

We hebben gezien dat er verband is tussen de grootte van onze steekproef en de onzekerheid op het bepaalde gemiddelde. Deze volgt de $\sqrt{n}$-wet. We gaan in deze opdracht een lineaire regressie (ofwel een fit) aan de data punten maken met behulp van de kleinste kwadraten methode.

We gaan eerst even terug naar het experiment om te kijken wat we ook alweer aan het bepalen waren. We hebben een ton met kogels en een heel nauwkeurige weegschaal. We kunnen ons verschillende vragen stellen over de massa van de kogels.

- Als we een kogel uit de ton pakken: "Wat is de massa van deze kogel?". De massa van een enkele kogel weten we in dit experiment met bijna oneindige precisie. In ons voorbeeld zelfs bijvoorbeeld: $m_{kogel} = 85.07426079254506 \pm 0.0000000000001$ gram.

- Wat is de massa van een ***typische*** kogel. Wat we hiermee bedoelen is: Als ik een *willekeurige* kogel uit de ton pak, wat is dan de massa? Het antwoord op deze vraag kun je vinden als je het gemiddelde weet van de kogels in de ton en de spreiding (standaarddeviatie) op de massa's. Stel dat het gemiddelde van de populatie 25.0 gram is en de standaarddeviatie 2.5 gram dan zeg je in dat geval dat een **typische** massa: $m_{\text{kogel}} = 25.0 \pm 2.5$ gram is. Je moet dan dus wel het gemiddelde en de spreiding weten, of bepalen. De standaarddeviatie is hier dus een maat voor de onzekerheid.
- Om de bovenstaande vraag te kunnen beantwoorden moet je dus weten wat het gemiddelde van de kogel massa's is en wat de spreiding op deze massa's is. De derde vraag die je kunt stellen is dus: Wat is het gemiddelde van de kogel massa's. Om die vraag te beantwoorden kunnen we een steekproef nemen uit de ton. We zien dan al snel een spreiding ontstaan. Bij de eerste kogel kunnen we nog heel weinig zeggen over het populatie gemiddelde en zeker niets over de spreiding. Bij twee kogels heb je al wat meer informatie. Mocht je de standaarddeviatie σ_m kennen dan kun je uitrekenen wat de onzekerheid is op het steekproef gemiddelde met de $\sqrt{n}$ wet. Als je steekproef redelijk groot is, dan kun je ook de spreiding s_n hiervoor gebruiken. De onzekerheid waar we hier over hebben gaat dus niet over de onzekerheid op de massa van een typische kogel, maar over de onzekerheid op de centrale waarde van het kogelmassagemiddelde zelf.

We willen dus weten wat de een **typische** kogel uit de ton weegt, we nemen een steekproef om het gemiddelde van de kogels in de ton te bepalen en we onderzoeken hoe de onzekerheid op de centrale waarde van dit gemiddelde afhangt van de grootte van de steekproef. We focussen dus op de spreiding van de bepaalde gemiddeldes. In M2.1 hebben we een lineair verband gezien tussen de spreiding van de bepaalde gemiddeldes en de grootte van de steekproef. In deze opdracht gaan we deze nu 'fitten' met behulp van de kleinste kwadraten methode..

We gaan eerst datapunten fitten met gelijke fouten. Later kijken we naar meer realistische onzekerheden op de datapunten. Met de volgende instructie kun je de datapunten opvragen:

```
inv_sqrt_n,std_n,std_n_err = ds.GroteAantallenFitSetGenerator()
```

De `inv_sqrt_n` punten zijn de waardes van $1/\sqrt{n}$ waarbij n de grootte is van de steekproef zoals je die in M2.1 hebt gedaan, `std_n` is de onzekerheid s_{g_n} en `std_n_err` zijn de onzekerheden op de waardes van s_{g_n}. In deze opdracht noteren we s_{g_n} als s_n en de onzekerheid op deze standaarddeviatie als Δs_n.

Voor deze dataset zijn de waardes van Δs_n dus nog allemaal gelijk, later in deze opdracht zullen we met meer realistische onzekerheden gaan werken. Maar eerst gaan we de fit opzetten.

Naar aanleiding van de $\sqrt{n}$-wet verwachten dat de relatie tussen n en s_n er als volgt uitziet:

$$s_n = \sigma/\sqrt{n}.$$

De parameter σ is nu de standaarddeviatie van de originele verdeling van de massa van de kogels, dus van de gehele populatie. De variabele $\hat{\sigma}$ is de geschatte waarde van σ die we proberen te vinden met de fit.

- Maak eerst een grafiek waarbij je `std_n` tegen `inv_sqrt_n` uitzet met de foutenvlaggen. Gebruik hier niet de code voor uit M2.1 maar maak gebruik van de dataset die je met het commando dat hier boven beschreven staat verkrijgt. Kijk goed naar de punten en probeer alvast voor jezelf in te schatten welke waarde je verwacht voor σ.
- Vind nu de meest optimale waarde van σ door gebruik te maken van de kleinste-kwadraten methode.

    1. Schrijf eerst een functie die voor een waarde van `inv_sqrt_n` en een gegeven waarde voor σ een waarde teruggeeft voor de voorspelling van `std_n`. Gebruik hierbij de formule die hierboven gegeven is.

    2. Schrijf een functie die de χ^2 uitrekent volgens de formule die je vindt in het hoofdstuk de kleinste-kwadraten.
    3. Schrijf een loop die over verschillende waardes van σ loopt voor het optimalisatie proces en voor elke waarde van σ de χ^2 uitrekent.

4. Vind nu voor welke waarde van σ de laagste waarde van χ^2 voorkomt. Dit is je schatting $\hat{\sigma}$.
   Tip: Weet je zeker dat de juiste waarde van σ in het gebied ligt waar je probeert te optimaliseren? Probeer met de grafiek die je eerder maakte af te schatten welke waarde voor σ je verwacht te vinden.

- **M3.1a) Welke waarde voor σ geeft de beste fit? Met andere woorden wat is, na het optimaliseren met de kleinste kwadraten methode, je geschatte $\hat{\sigma}$?**

- **M3.1b) Maak een grafiek met de datapunten, de foutenvlaggen en het fit resultaat.**
  Tip: De gefitte functie kun je het makkelijkste plotten door met behulp van de `inv_sqrt_n` lijst een bijbehorende lijst te maken met behulp van de functie die je in stap 1 hebt gemaakt.

- **M3.1c) Maak een grafiek waarin je de waarde voor χ^2 uitzet tegen σ. Bij welke waarde van χ^2 vind je $\hat{\sigma}$?**

Met de functie:

```
s_true = ds.GroteAantallenStdTrue()
```

Kun je de werkelijke 'true'-waarde van σ terugvragen.

- **M3.1d) Controleer of jouw gefitte waarde van $\hat{\sigma}$ overeen komt met je uitkomst met je uitkomst voor `s_true`. Je verwacht altijd nog wel wat verschillen te zien - vooral omdat de onzekerheden op de waardes van `s_n` niet realistisch waren.**

We gaan nu de fit uitvoeren met realistische onzekerheden op de datapunten. Deze datapunten genereer je met de volgende functie:

```
inv_sqrt_n,std_n,std_n_err = ds.GroteAantallenStdGenerator()
```

- **M3.1e) Vind nu de meest optimale waarde van $\hat{\sigma}$ door gebruik te maken van de realistische foutenvlaggen. Bij welke χ^2 ligt deze optimale waarde?**

- **M3.1f) Maak nu een grafiek met de datapunten, de foutenvlaggen en het fit resultaat voor de dataset met reële foutenvlaggen.**

- **M3.1g) Vergelijk nu de gevonden $\hat{\sigma}$ met de 'true' waarde van σ. Komt deze nu meer of minder overeen in vergelijking met je eerste fit?**
- **M3.1h) Bereken nu de gereduceerde χ^2, dat wil zeggen corrigeer de gevonden χ^2 voor het aantal vrijheidsgraden van de fit. Interpreteer nu deze χ^2/df. Is deze beter of slechter dan een $\chi^2/df = 0.1$? Zoals gebruikelijk, beredeneer je antwoord.**

5.2 M3.2 Halfwaardedikte III ***

In opgave M2.3 hebben we gezien dat de meetmethode die we gebruikten om de halfwaardedikte te bepalen niet optimaal was. Er was zeker sprake van een onzuivere meting doordat we stelselmatig een te hoge waarde van d_{half} terugkregen.

In deze opgave zullen we zien dat de onzuiverheid te maken heeft met de methode waarop we de halfwaardedikte hebben bepaald. Het heeft niets te maken met de opstelling van

de meting of met de verzamelde datapunten. Het is de analyse techniek die zorgt voor de onzuiverheid.

In deze opdracht gaan we een fit gebruiken om de waarde van d_{half} te achterhalen. In opdracht M3.1 hebben we onze eigen lineaire regressie methode geprogrammeerd met behulp van de kleinste kwadraten methode. In deze opdracht gebruiken we een fit pakket `lmfit`. Dit programma rekent de χ^2 uit en minimaliseert deze voor ons. Dat scheelt op zich een hoop werk, maar je zal in deze opdracht zien dat het toch ook weer niet helemaal vanzelf gaat.

Om dit fit pakket te kunnen gebruiken moet het volgende import statement gebruiken:

```
from lmfit import models
```

Maak nu een dataset aan met de standaard waardes zoals je dat in M2.3 ook hebt gedaan:

```
counts, diktes, dtrue = ds.DataSetHalfwaardeDikteVariatie()
```

We hebben eerst een functie nodig die de datapunten beschrijft en een set met startwaardes voor de parameters. We kijken eerst nog even naar de formule die in M1.1 is gegeven:

$$I(d; N_0, d_{half}) = I_0 \times \left(\frac{1}{2}\right)^{d/d_{half}} \tag{5.1}$$

We zien dat de functie uiteindelijk afhangt van twee parameters: I_0 en d_{half}. De waardes I_0 en $I(d)$ zijn natuurlijk direct gerelateerd aan de gemeten waardes voor N_0 en $N(d)$. In principe is N_0 en dus I_0 gemeten, maar hier zit een (bekende) onzekerheid op en om die reden wil je hem 'vrijlaten' in de fit.

De functie `functie` die we straks gebruiken voor de fit ziet er in het algemeen als volgt uit:

```
def functie(x,par1,par2) :
    y = een formule
    return y
```

De parameters die hier worden meegegeven zijn de parameters die worden geschat (of geoptimaliseerd) in de fit. Voor deze twee waardes zullen we straks ook de startwaardes moeten meegeven.

1. Schrijf nu eerst de code voor de functie `functie(d, N0, dhalf)` die de relatie tussen dikte d en de counts aangeeft. Controleer of die goed werkt.
2. Voor de fit hebben we ook een lijst met gewichten nodig, noem deze `N_inv_err`. Dit zijn de reciproke waardes van de fouten op de counts. Maak hiervoor een lijst aan. Als de onzekerheid op N, ΔN is, dan is het gewicht $1/\Delta N$.

Als we onze functie en de lijst met gewichten hebben gedefinieerd dan kunnen we de fit uitvoeren.

```
ons_model = models.Model(functie)
result= ons_model.fit(counts, d=diktes, weights =N_inv_err, N0=startwaarde, dhalf=startwaarde)
```

We definiëren eerst `ons_model` en vervolgens fitten we deze. Je moet een aantal opties meegeven:

```
result   : deze vangt het fit resultaat op
counts   : de lijst met counts
d=diktes : d is de eerste parameter van functie, diktes is de lijst met diktes
weights = N_inv_err : hier geef je de lijst met gewichten mee
N0= startwaarde : hier moet je de startwaarde voor de fit meegeven op N0
dhalf = startwaarde : hier moet je de startwaarde voor dhalf meegeven
```

Je ziet dat je nog zelf twee startwaardes mee moet geven voordat de fit kan werken.
Met het volgende commando kun je de fitresultaten uitprinten:

```
print(result.fit_report())
```

- **M3.2a) Voer de fit uit en bekijk het resultaat. Als je tevreden bent met de fit kopieer dan je resultaat op het inlevertemplate. Het kan zijn dat je de startwaardes van de parameters nog iets moet aanpassen als de fit niet convergeert.**

De gefitte curve kunnen we ook weergeven in een grafiek. Maak zoals gebruikelijk een grafiek met foutenvlaggen. Het fitresultaat kun je dan als volgt toevoegen:

```
plt.plot(diktes, result.init_fit, 'k--', label='initial fit')
plt.plot(diktes, result.best_fit, 'r-', label='best fit')
plt.legend(loc='best')
```

- **M3.2b) Maak een grafiek met de datapunten, foutenvlaggen en het gefitte resultaat. Maak de grafiek netjes af.**

- **M3.2c) Bekijk de gereduceerde χ^2/df. Ziet deze waarde er goed uit? Beredeneer je antwoord. Wat is het aantal vrijheidsgraden in de fit?**

- **M3.2d) Wat is de geschatte waarde $\hat{d}_{half}$? Vergelijk deze met de 'true' waarde 'dtrue'.**

- **M3.2e) De correlatiecoëfficiënt ρ wordt ook uitgeprint. Hoe groot is deze en wat zegt dat?**

Definieer nu een polynoom met de volgende code:

```
def poly(d,N0,a,b) :
    y = N0 + a*d + b*d*d
    return y
```

Fit deze functie aan de datapunten, zorg dat de startwaardes zo worden ingesteld dat de fit convergeert.

- **M3.2f) Maak een grafiek met de datapunten, foutenvlaggen en het gefitte resultaat. Maak de grafiek netjes af.**

- **M3.2g) Presenteer de fitresultaten van de poly fit op het inlevertemplate.**

- **M3.2h) Vergelijk nu de twee fits met elkaar. Bekijk de uitkomsten van de gefitte exponentiele functie met de gefitte polynoom. Welke functie beschrijft de data het beste? Op basis van welke variabelen trek je deze conclusie? Beargumenteer je antwoord.**