Data Analyse en Statistiek

Hella Snoek & Marthe Schut

Inhoud

Hoofdstuk 1

Foutenpropagatie

Vaak kunnen we de grootheid die we willen weten niet direct meten, maar meten we een observabele die zich via een bepaalde functie verhoudt tot de gezochte grootheid. Of meten we zelfs twee of meer variabelen die we nodig hebben om de gewilde grootheid te bepalen.

Dit is bijvoorbeeld het geval als we de gemiddelde snelheid van een auto willen bepalen. Dit zouden we kunnen doen door de tijd te meten die de auto nodig heeft om een bepaald traject af te leggen. We meten dan de door de auto gebruikte tijd, T en de lengte van het traject, L, en die zetten we dan om in snelheid via de bekende formule $v = L/T$. Of we bepalen bijvoorbeeld de massa van een elementair deeltje (in rust) en willen dit omzetten naar de energie van het deeltje via de formule $E = mc^2$.

Als we de onzekerheid weten op de gemeten grootheden dan kunnen we deze omzetten naar de grootheid die we eigenlijk willen bepalen. Dit noemen we het propageren van fouten. In dit hoofdstuk leren we je de basisregels voor het propageren van **ongecorreleerde** fouten. Dat wil zeggen dat als er meerdere onzekerheden worden gepropageerd deze onzekerheden onafhankelijk zijn; De meting van de ene observabele heeft geen invloed op de meting van de andere observabele; de fout die we maken in het meten van de ene grootheid hangt niet af van de fout die we maken op de andere gemeten grootheid.

Het is goed om alvast te beseffen dat er ook gecorreleerde fouten bestaan. Er zijn twee oorzaken voor het ontstaan van gecorreleerde fouten:

- Doordat er in de meting een correlatie is. Een voorbeeld van een gecorreleerde fout is als we een oppervlakte van een tafel willen weten en we meten de lengte en de breedte met hetzelfde meetlint op. Als het meetlint een afwijking heeft waardoor we de lengte te groot opmeten, dan zullen we waarschijnlijk ook de breedte te groot opmeten.
- Doordat er een onderliggende parameter is waar beide gemeten grootheden vanaf hangen.

Hier behandelen we dus alleen ongecorreleerde fouten.

1.1 Basisregel

We beginnen met de **algemene regel voor het propageren van ongecorreleerde fouten**. Daarna zullen we laten zien hoe deze regel eruitziet voor eenvoudige relaties. Deze zou je apart kunnen leren, maar je kunt ook altijd de basisregel gebruiken. Het resultaat behoort hetzelfde te zijn. We noteren de onzekerheid op variabele x in dit hoofdstuk met Δx waar we eerder ook wel σ_x hebben gebruikt.

Als $q = q(x, y, z, \dots)$ een functie is met meerdere ongecorreleerde variabelen, dan wordt de onzekerheid op q gegeven door:

$$\Delta q = \sqrt{\left(\frac{\delta q}{\delta x}\Delta x\right)^2 + \left(\frac{\delta q}{\delta y}\Delta y\right)^2 + \left(\frac{\delta q}{\delta z}\Delta z\right)^2 + \dots} \tag{1.1}$$

Hierbij zijn $\frac{\delta q}{\delta x}$, $\frac{\delta q}{\delta y}$ etc. de partiële afgeleiden van q naar de betreffende variabele.

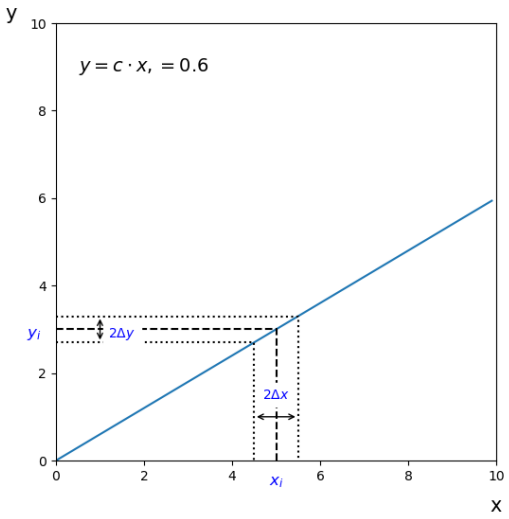
We zullen laten zien hoe deze formule werkt aan de hand van een paar voorbeelden.

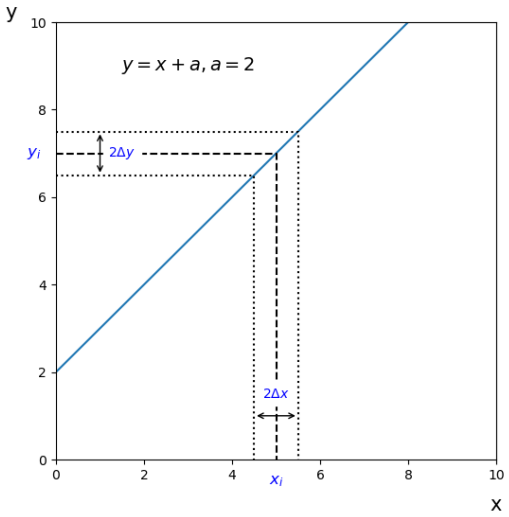
Voorbeeld 1: Factor
Stel we hebben een vergelijking $y = c \cdot x$ met een standaarddeviatie op x van Δx. Dan is de standaarddeviatie op y, (Δy), gelijk aan:

$$\Delta y = \sqrt{\left(\frac{\delta y}{\delta x}\Delta x\right)^2} = c \cdot \Delta x. \tag{1.2}$$

In dit geval schaalt de onzekerheid op x (Δx) dus met dezelfde factor c tot de onzekerheid op y (Δy). In het plaatje hieronder wordt voor een willekeurige waarde x_i het effect van de propagatie van Δx rond de waarde x_i naar de fout Δy rond y_i visueel weergegeven. Je kunt duidelijk zien dat de grootte van Δy veranderd is met de factor c.

Voorbeeld 2: Translatie
Stel we hebben een vergelijking $y = x + a$ met een standaarddeviatie op x van Δx. Dan is de standaarddeviatie op y, (Δy), gelijk aan:
$\Delta y = \sqrt{\left(\frac{\delta y}{\delta x}\Delta x\right)^2} = \Delta x.$
Wederom geven we het effect van de foutenpropagatie van Δx rond x_i naar Δy rond y_i grafisch weer in het plaatje hieronder. Je ziet dat de translatie geen effect heeft op de grootte van de onzekerheid.

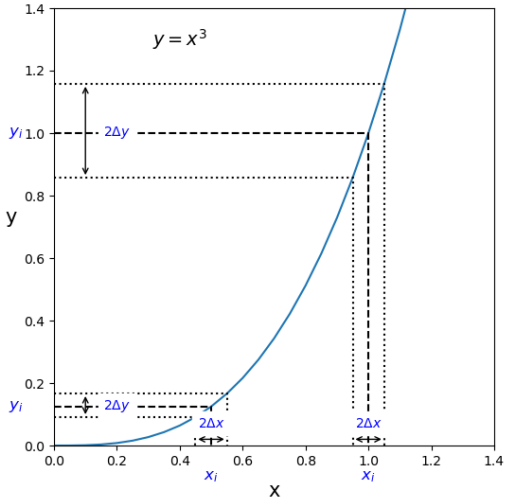




Voorbeeld 3: Macht
Stel we hebben een vergelijking $y = x^3$ met een standaarddeviatie op x van Δx. Dan is de standaarddeviatie op y, (Δy), gelijk aan:

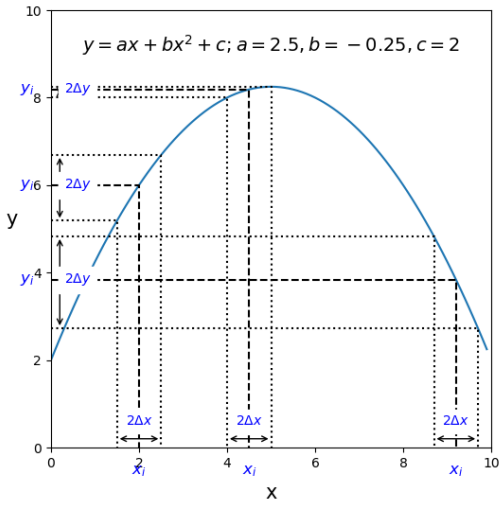
$\Delta y = \sqrt{\left(\frac{\delta y}{\delta x}\Delta x\right)^2} = 3x^2 \cdot \Delta x.$

Het effect van de foutenpropagatie volgens deze formule van Δx rond x_i naar Δy rond y_i wordt weer grafisch weergegeven in het plaatje hieronder. Je kunt zien dat de mate waarin de grootte van Δx verandert afhangt van de gekozen waarde van x_i, op sommige plekken is hij kleiner geworden, op andere plekke groter.



Voorbeeld 4
Stel we hebben een vergelijking $y = ax + bx^2 + c$ met een standaarddeviatie op x van Δx. Dan is de standaarddeviatie op y, (Δy), gelijk aan:
$\Delta y = \sqrt{\left(\frac{\delta y}{\delta x}\Delta x\right)^2} = (a + 2bx)\Delta x.$
In het plaatje hieronder geven we nu voor verschillende waardes x_i de foutenpropagatie van Δx naar Δy de grafische interpretatie. We zien dat het niet alleen de relatieve grootte van Δy afhangt van de gekozen waarde van x_i maar dat op sommige plaatsen de boven en ondergrens van de onzekerheid zijn geïnverteerd.

Voorbeeld 5
Stel we hebben een vergelijking $z = ax + y^2$ met standaarddeviaties Δx en Δy . Dan is de standaarddeviatie op z, (Δz), gelijk aan:
$\Delta z = \sqrt{\left(\frac{\delta z}{\delta x}\Delta x\right)^2 + \left(\frac{\delta z}{\delta y}\Delta y\right)^2} = \sqrt{(a\Delta x)^2 + (2y\Delta y)^2}.$
Voorbeeld 6
Stel we hebben een vergelijking $z = ax + y^2 + 2xy$ met standaarddeviaties Δx en Δy . Dan is de standaarddeviatie op z, (Δz), gelijk aan:



$$\Delta z = \sqrt{\left(\frac{\delta z}{\delta x}\Delta x\right)^2 + \left(\frac{\delta z}{\delta y}\Delta y\right)^2} = \sqrt{((a+2y)\cdot\Delta x)^2 + ((2y+2x)\cdot\Delta y)^2}.$$

1.2 Som en verschil

De algemene regel kan eenvoudig worden uitgeschreven naar de regel voor som en verschil. Als $q = x + y$ of $q = x - y$ dan wordt de onzekerheid op q gegeven door:

$$\Delta q = \sqrt{\left(\frac{\delta q}{\delta x}\Delta x\right)^2 + \left(\frac{\delta q}{\delta y}\Delta y\right)^2} = \sqrt{(\Delta x)^2 + (\Delta y)^2}. \tag{1.3}$$

We mogen de varianties $(\Delta x)^2$ en $(\Delta y)^2$ in het geval van een vergelijking met enkel sommen en/of verschillen dus optellen.

1.3 Vermenigvuldigen met constante

Als q een exacte veelvoud c is van de gemeten waarde x, dus $q = c \cdot x$, dan geldt:

$$\Delta q = \sqrt{\left(\frac{\delta q}{\delta x}\Delta x\right)^2} = |c|\Delta x. \tag{1.4}$$

De onzekerheid op q is dus gelijk aan de onzekerheid op x geschaald met dezelfde factor c.

1.4 Vermenigvuldigen met variabelen

Als q een vermenigvuldiging is van meerdere variabelen, dus bijvoorbeeld $q = x \cdot y \cdot z$ dan geldt:

$$\Delta q = \sqrt{\left(\frac{\delta q}{\delta x}\Delta x\right)^2 + \left(\frac{\delta q}{\delta y}\Delta y\right)^2 + \left(\frac{\delta q}{\delta z}\Delta z\right)^2} = \sqrt{\left(\frac{q}{x}\Delta x\right)^2 + \left(\frac{q}{y}\Delta y\right)^2 + \left(\frac{q}{z}\Delta z\right)^2}. \quad (1.5)$$

Dit kan je eenvoudiger schrijven als:

$$\frac{\Delta q}{q} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 + \left(\frac{\Delta z}{z}\right)^2 + \left(\frac{\Delta z}{z}\right)^2}. \quad (1.6)$$

Ofwel de relatieve fout $\frac{\Delta q}{q}$ is gelijk aan de kwadratische som van de variabelen.

Voorbeeld - foutenpropagatie en afronding van de getallen
Stel dat we de lengte van het blokje hebben gemeten en we lezen de volgende waarde af:

- De lengte (l) $= 7.60 \pm 0.10$ cm
- De breedte (b) $= 4.10 \pm 0.20$ cm
- De hoogte (h) $= 2.00 \pm 0.20$ cm

Het volume van het blokje wordt gegeven door:
$V = l \cdot b \cdot h = 7.60 \cdot 4.10 \cdot 2.00 = 62.32$ cm^3
We gebruiken de regel dat als $q = x \cdot y \cdot \ldots$ dan:
$\frac{\Delta q}{|q|} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 \left(\frac{\Delta y}{y}\right)^2 + \left(\frac{\Delta z}{z}\right)^2}$
Dus:

$$\begin{aligned}\frac{\Delta V}{|V|} &= \sqrt{\left(\frac{\Delta l}{l}\right)^2 + \left(\frac{\Delta b}{b}\right)^2 + \left(\frac{\Delta h}{h}\right)^2} \\ &= \sqrt{\left(\frac{0.1}{7.6}\right)^2 + \left(\frac{0.2}{4.1}\right)^2 + \left(\frac{0.2}{2.0}\right)^2} \\ &= 0.01255\ldots\end{aligned}$$

We ronden dit nog niet af, dat doen we pas als we de absolute fout hebben:

$$\begin{aligned}\Delta V &= \frac{\Delta V}{|V|} \cdot |V| \\ &= 0.01255 \cdot \cdots \cdot 62.32 \\ &= 0.78228\ldots \\ &\approx 0.78\end{aligned}$$

Het gemeten volume van het blokje is dus $V = 62.32 \pm 0.78$ cm^3

Hoofdstuk 2

Wet van Grote Aantallen

In opgave M1.4 hebben we gezien hoe de spreiding van een gemeten gemiddelde van metingen steeds kleiner wordt als we meer data gebruiken om het gemiddelde te bepalen. Dit is een belangrijke observatie. Het geeft aan dat hoe meer data we hebben, hoe nauwkeuriger we ons resultaat weten. Je voelt misschien al aan dat dit niet altijd op gaat. Wanneer dit wel en wanneer dit niet opgaat zullen we hier bespreken.

We bespreken hier twee regels, of wetten, de $\sqrt{n}$-wet en de wet van grote aantallen. De eerste wet zegt dat we een gemiddelde, onder bepaalde voorwaarden, steeds beter kennen als we meer datapunten meenemen. De tweede wet zegt dat het gemiddelde van de steekproef langzaam zal convergeren naar het gemiddelde van de populatie.

2.1 De $\sqrt{n}$-wet

We kijken naar twee onafhankelijke stochasten, X en Y. De verwachtingswaarde van $X+Y$ is gelijk aan:

$$E(X+Y) = E(X) + E(Y) \tag{2.1}$$

Als X en Y onafhankelijk zijn dan geldt ook:

$$Var(X+Y) = Var(X) + Var(Y) \tag{2.2}$$

Het ziet er misschien ingewikkeld uit, maar het enige wat we doen is een nieuwe variabele definiëren die de som is van twee variabelen. De variantie op de som vinden we via de gewone fouten propagatie regels.

Stel nu dat we dit uitbreiden. En we nemen de som van N onafhankelijk stochasten, $X_1, X_2, ..., X_N$ die elk dezelfde onderliggende verdeling kennen. Dat wil zeggen dat ze allemaal dezelfde verwachtingswaarde en dezelfde variantie hebben.

NB. De verwachtingswaarde is niet gelijk aan de gemeten waarde. Kijk voor dit verschil nog eens naar basisbegrippen in module 1. Als je terugdenkt aan de opgave van de kogels is de verwachtingswaarde van de massa van een kogel gelijk aan de gemiddelde massa van alle kogels. Als we een willekeurige kogel uit de ton pakken, dan is de gemiddelde massa van de kogels in de ton, de *verwachting* die we hebben van de massa van de kogel die we pakken. De verwachtingswaarde is dus hier het gemiddelde.

De variantie is de spreiding op de massa distributie. Als je een willekeurige kogel uit de ton pakt is de kans heel klein dat de massa precies gelijk is aan de verwachtingswaarde van de massa. De variantie geeft aan in welk gebied van waardes we verwachten de massa van de kogel te vinden.

De formule voor de som kunnen we nu schrijven als:

$$Som_N = X_1 + X_2 + ... + X_N. \tag{2.3}$$

En het gemiddelde kunnen we schrijven als:

$$E(< X_1...X_N >) = \frac{Som_n}{N}. \tag{2.4}$$

Als de verwachtingswaarde van een enkele stochast $E(X_i)$ gelijk is aan het gemiddelde μ en de variantie gelijk is aan $Var(X_i) = \sigma^2$, dan geldt nu voor de verwachtingswaarde van de som:

$$E(S_N) = \mu N \tag{2.5}$$

en voor het gemiddelde:

$$E(< X_1...X_N >) = \mu. \tag{2.6}$$

En dan geldt voor de variantie

$$Var(S_N) = N\sigma^2 \tag{2.7}$$

en

$$Var(< X_1...X_N >) = \frac{\sigma^2}{N}. \tag{2.8}$$

Dit betekent dat **de standaarddeviatie van de som van de stochasten** gelijk is aan $\sigma \cdot \sqrt{N}$.
De standaarddeviatie van het gemiddelde is dan gelijk aan:

$$\text{standaarddeviatie op het gemiddelde is: } \frac{\sigma \cdot \sqrt{N}}{N} = \frac{\sigma}{\sqrt{N}}. \tag{2.9}$$

Dit betekent dus dat als we het gemiddelde van de massa van N aantal kogels nemen waarbij de kogels een Normale distributie hebben met een gemiddelde μ en een standaarddeviatie van σ, de onzekerheid op de bepaalde gemiddelde massa gelijk is aan $\sigma/\sqrt{N}$.
Hoe meer kogels we wegen en meenemen in ons gemiddelde, hoe nauwkeuriger we dit gemiddelde kennen.

2.2 De wet van Grote Aantallen

Intuïtief voelen we aan dat hoe meer metingen we doen, hoe meer informatie we hebben, en hoe nauwkeuriger ons resultaat is.

De **wet van grote aantallen** zegt dat het berekende steekproef gemiddelde, $< X >$, van een distributie met een eindige variantie, convergeert naar het populatie gemiddelde μ voor steeds grote steekproeven:

$lim_{N\to\infty} P(| < X > -\mu| > \epsilon) = 0$

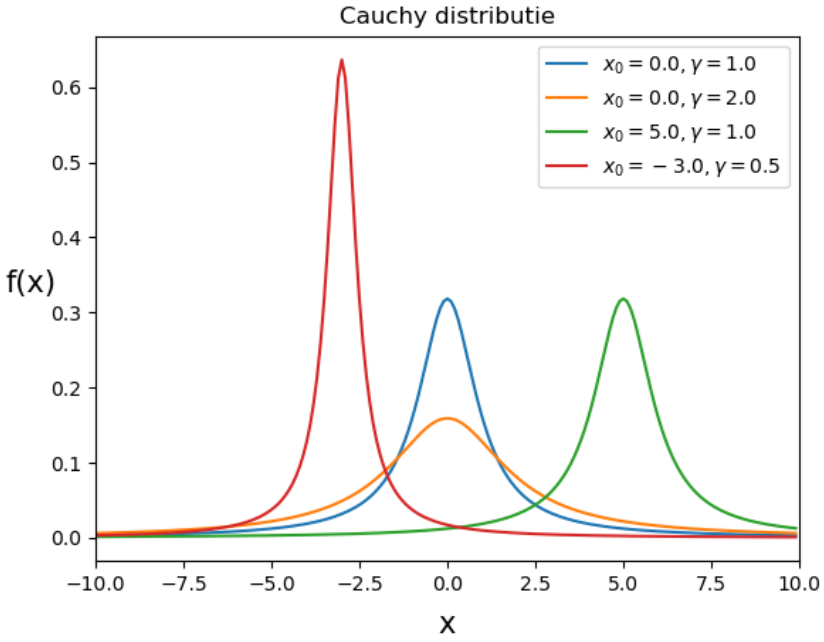
Ofwel de kans dat het steekproef gemiddelde meer afwijkt van het populatie gemiddelde dan een heel klein getal convergeert naar 0 voor oneindig grote steekproeven. Voor eindige populaties is dit natuurlijk zeker waar. Maar denk hier ook aan oneindig grote, of nagenoeg oneindig grote populaties, zoals bijvoorbeeld als je de gemiddelde massa van het electron wilt bepalen.

Tip: In deze video wordt de wet van grote aantallen nogmaals duidelijk uitgelegd.

Als je de wet goed leest zie je dat er een voorwaarde aan vast zit. Namelijk dat de variantie van de stochast eindig moet zijn, en dat dus de verwachtingswaarde van de stochast bepaald is. Er bestaan distributies, zoals de Cauchy of de Landau distributie waarvoor dit dus niet geldt. Deze distributies hebben oneindig lange staarten. Hieronder zie je hoe de Cauchy distributie eruit ziet.

Wiskundig kan de wet van de grote aantallen dus weleens voor problemen zorgen. In Natuurkundige experimenten zijn verdelingen uiteindelijk vaak beknot door bijvoorbeeld de eindigheid van energie. Voor Natuurkundige experimenten gaat de wet van grote aantallen dus vaak wel op.

Overigens noemen we deze wet van grote aantallen de *zwakke* wet van grote aantallen, er bestaat ook een *sterke* wet. We gaan hier niet in op de kleine verschillen tussen deze twee wetten, online kun je er eventueel genoeg over vinden.


Cauchy distributie
$x_0 = 0.0, \gamma = 1.0$
$x_0 = 0.0, \gamma = 2.0$
$x_0 = 5.0, \gamma = 1.0$
$x_0 = -3.0, \gamma = 0.5$
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
x

Hoofdstuk 3

Meerdimensionale datasets

Het komt vaak voor dat we datasets hebben waarbij we meerdere variabelen tegelijkertijd hebben gemeten. Bijvoorbeeld als we een steekproef doen onder de bevolking waarbij we allerlei gegevens tegelijkertijd opvragen zoals leeftijd, inkomen, gezinssamenstelling etc. We kunnen dan niet alleen naar verdelingen kijken van bijvoorbeeld alleen het inkomen, maar we kunnen ook naar het inkomen kijken *afhankelijk* van de leeftijd. Dit levert dus meer informatie op dan dat we deze gegevens afzonderlijk zouden hebben verzameld. Ook in natuurkundige experimenten komen multidimensionale datasets veel voor.

Voor elke afzonderlijke variabele kunnen we bijvoorbeeld het gemiddelde en de standaarddeviatie berekenen met behulp van de formules die we in 'Basisbegrippen' hebben geïntroduceerd. Maar we kunnen nu ook kijken of de waarde van een observabele afhangt van een andere observabele in de dataset. Dit noemen we correlatie. Ook kunnen we berekenen of een spreiding in een variabele afhangt van de waarde van een andere variabele. We noemen die covariantie. Hieronder introduceren we eerst covariantie, daaronder komt correlatie aan bod.

3.1 Variantie en covariantie

De variantie geeft zoals eerder besproken (onder 'Basisbegrippen') een maat voor de spreiding van een dataset aan. Bij een 2D dataset waarbij een variabele wordt aangegeven op de x-as en een andere variabelen op de y-as wordt de mate van spreiding o.a. aangegeven met de *covariantie*.

De covariantie bij een 2D dataset geeft aan in welke mate de data verspreid is over het twee dimensionale vlak.

Voor twee variabelen x en y wordt de covariantie aangeduid met $cov(x, y)$ en gegeven door:

$$cov(x, y) = E((x - E_x)(y - E_y)) \tag{3.1}$$

Hier staat E voor de *verwachtingswaarde*. De verwachtingswaarde voor

$$x \tag{3.2}$$

en y worden respectievelijk aangegeven met E_x en E_y. De formule geeft dus aan dat de covariantie gelijk is aan de verwachtingswaarde van het verschil tussen de waarde van de variabele x en de verwachtingswaarde van x vermenigvuldigd met het verschil tussen de variabele y en de verwachtingswaarde van y.

Als voor waardes x die bovengemiddeld zijn, overwegend samen gaan met relatief hoge waardes van y, dan hebben we te maken met een positieve waarde voor de covariantie. Als bij de waarden voor relatief hoge waardes van x de waardes van y voornamelijk onder de verwachtingswaarde liggen, dan is de covariantie negatief. Als de covariantie gelijk is aan nul dan is er, gemiddeld over de hele dataset, geen afhankelijkheid. Het kan zijn dat voor delen van de dataset wel degelijk een positieve covariantie bestaat, deze wordt dan opgeheven door een ander gedeelte met een negatieve covariantie.

Met de volgende formules kun je de covariantie van een dataset uitrekenen:

- Voor discrete verdelingen geldt :

$$cov(x, y) = \frac{1}{n} \sum_{i}^{n} (x_i - < x >) \cdot (y_i - < y >). \tag{3.3}$$

- Voor continue verdelingen geldt:

$$cov(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f(x, y) dy dx \tag{3.4}$$

De covariantie geeft dus aan in hoeverre waarden van de ene variabele toenemen/afnemen bij toenemende waarden van de andere variabele. De covariantie is een heel nuttige maat maar lastig te interpreteren vanwege de dimensies die, net als bij de variantie, niet dezelfde zijn als de variabelen zelf. Eenvoudiger is om naar de correlatiecoëfficiënt ρ te kijken.

Hierkun je een filmpje zien die covariantie uitlegt.

3.2 Correlatie

De correlatiecoëfficiënt ρ is gedefinieerd als:

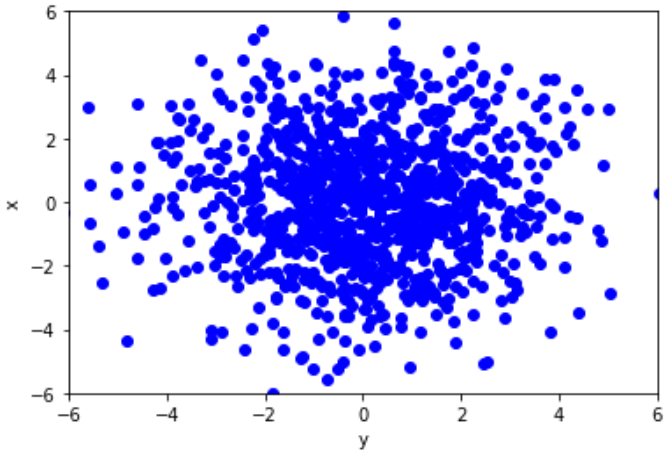
$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y} \tag{3.5}$$

Hierbij is $cov(x, y)$ de covariantie tussen variabele x en variabele y, en zijn σ_x en σ_y de standaardafwijkingen van variabele x en y respectievelijk. Deze reken je dus uit met de formule die hierboven is gedefinieerd.

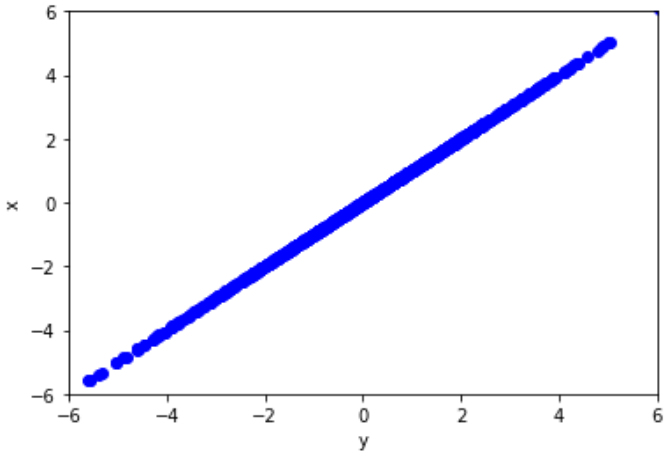
Als er geen correlatie is tussen de twee variabelen, dan is correlatiecoëfficiënt gelijk aan nul. Is de correlatiecoëfficiënt tussen de twee variabelen gelijk aan 1 of aan -1 dan zijn de twee variabelen maximaal afhankelijk. In het geval van een correlatiecoëfficiënt gelijk aan 1 is dit een positief lineair verband, in het geval van een correlatiecoëfficiënt gelijk aan -1 is dit een lineair verband met negatieve helling.

Hieronder zijn een aantal 2D datasets weergegeven met verschillende correlatiecoëfficiënten:

Dataset met een correlatiecoëfficiënt $\rho_{x,y} = 0$:



Dataset met een correlatiecoëfficiënt $\rho_{x,y} = 1$:

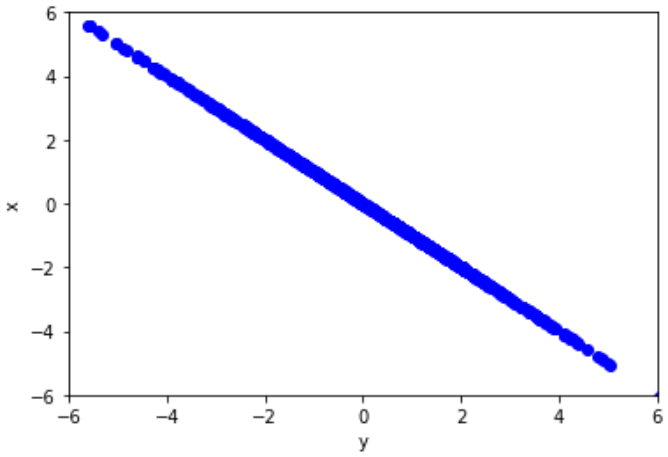


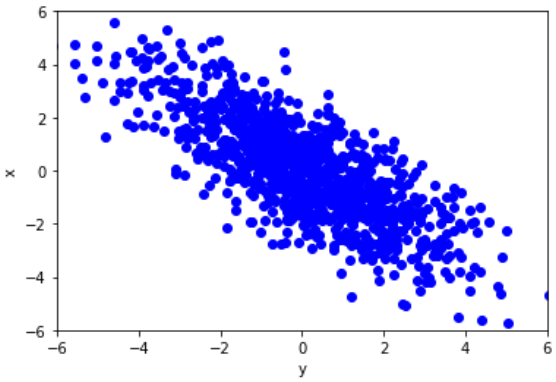
Dataset met een correlatiecoëfficiënt $\rho_{x,y} = -1$:

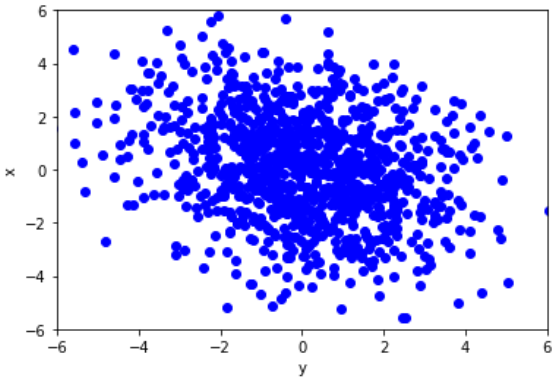
Datasets met een correlatiecoëfficiënt $\rho_{x,y} = -0.8$ en $\rho_{x,y} = 0.8$:

Datasets met een correlatiecoëfficiënt $\rho_{x,y} = -0.3$ en $\rho_{x,y} = 0.3$:

Hoe dichter de correlatiecoëfficiënt bij een waarde van 1 of -1 zit des te groter is de afhankelijkheid van de variabelen. Hoe te dichter de correlatiecoëfficiënt bij nul zit des te







kleiner is de correlatie tussen de variabelen.

Je kunt hiereen filmpje vinden waarin correlatie ook wordt uitgelegd. Er zijn meerdere 'spelletjes' op internet waarbij je kunt oefenen met het herkennen en raden van de correlatiecoëfficiënt van twee variabelen. Kijk bijvoorbeeld eens bij Geogebra-Correlatie game of Guess the correlation.

3.3 Correlatie en causaliteit

Soms betekent correlatie dat er oorzakelijk verband is tussen de twee observabelen. Dat wil zeggen dat de ene observabele invloed heeft op de andere observabele.

Een voorbeeld hiervan is bijvoorbeeld als je kijkt naar de ijsverkoop en de buitentemperatuur. Omdat het warm is buiten hebben mensen meer trek in een ijsje. Het is dus niet zo gek dat je er een verband tussen vindt. Dit verband noemen we een **causaal** verband. Iets wordt veroorzaakt door iets anders.

In wetenschappelijk onderzoek zijn we altijd op zoek naar correlaties. Immers, die kunnen wijzen op onbekende wetten of onderliggende, nog onbekende fenomenen. Toch moet je behoorlijk oppassen om meteen een conclusie te trekken. Niet alle observabelen die een gecorreleerd zijn hebben een causaal verband. Het kan ook toeval zijn, als je maar genoeg variabelen tegen elkaar uitzet zal je er altijd wat vinden die toevallig een correlatie vertonen. Het kan ook komen door een verborgen parameter. Dit wordt ook wel Simpsons paradox genoemd.

Een bekend voorbeeld van een Simpsons paradox is een onderzoek naar veiligheid op de scheepvaart. Er is gebleken dat er een positieve correlatie is tussen het dragen van reddingsvesten en het aantal ongevallen waarbij mensen verdronken zijn. Dit is natuurlijk niet wat je verwacht! Voordat je adviseert om alle reddingsvesten weg te laten gooien is het goed om nog iets verder onderzoek te plegen. Wat blijkt, de reddingsvesten worden alleen aangetrokken bij slecht weer op zee. De verborgen parameter is dus het weer. Als we de data nog een keer goed bekijken en nu kijken naar alleen de categorie slecht weer dan zien we dat de overlevingskans juist vele malen hoger als een reddingsvest wordt gedragen.

De les die je hieruit moet leren is dat je altijd heel goed moet nadenken over wat een verborgen parameter zou kunnen zijn en niet zomaar de conclusie trekken dat een correlatie ook causaliteit impliceert. Het is goed om zo'n conclusie eerst te onderbouwen met een plausibele verklaring.

Hoofdstuk 4

Extra kans rekenregels

In module 1 hebben we de complement, de en-regel en de of-regel geleerd voor het rekenen met kansen. Aan deze regels waren enkele voorwaarden verbonden.

De of-regel geldt alleen als de metingen A en B wederzijds uitsluitend zijn. Dat betekent dat een meting A niet kan voorkomen als B gemeten is.

> Een voorbeeld van kansen die niet wederzijds uitsluitend zijn is, als we weer kijken naar een set kaarten waar A bijvoorbeeld de kleur rood is en B het getal 4. Er bestaan rode kaarten met getal vier en in dit geval mogen we de kansen dus niet optellen.
>
> $P(\text{rood of } 4) \neq P(\text{rood}) + P(4)$

We breiden de regels hier verder uit en gaan kijken naar het combineren van kansen die niet wederzijds uitsluitend zijn. We kijken ook naar het begrip conditionele kans en introduceren Bayes theorema die gebruikt kan worden om informatie van kansen om te rekenen.

4.0.1 De of regel wanneer A en B niet wederzijds uitsluitend zijn:

In het geval A en B niet wederzijds uitsluitend zijn dan:

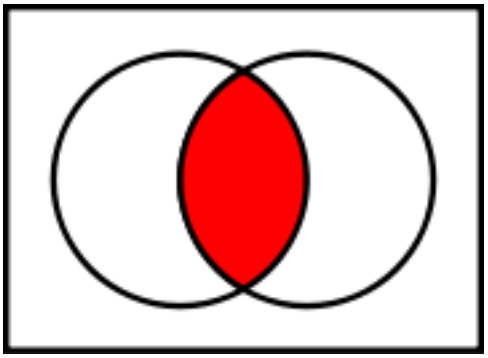
$P(\text{A en B}) \equiv P(A \cap B) > 0.$

De kans dat A of B gemeten wordt is dan:

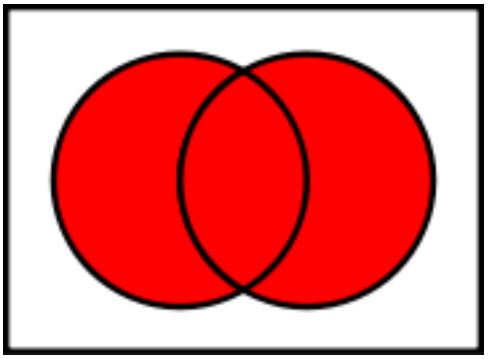
$P(\text{A of B}) = P(A) + P(B) - P(\text{A en B}).$

Voorbeeld: De kans dat een kaart rood is en een vier heeft is 2/52. De kans dat een kaart rood is of een vier is nu gelijk aan P(1/2) + P(4/52) - P(2/52) = 28/52.

De term (A en B) noemen we ook wel de doorsnede, of intersectie, van A en B. Het is het overlappende deel van elementen in de verzameling. Hieronder zie je het uitgebeeld in een Venn diagram. De doorsnede wordt ook wel genoteerd met $A \cap B$.
"doorsnede van A en B (bron wikipedia)"



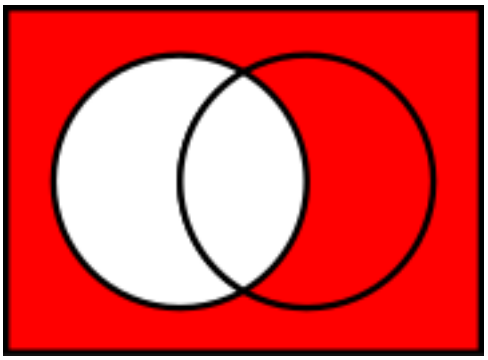
De vereniging van A en B wordt genoteerd met $A \cup B$ en is de verzameling van alle elementen van A en B. Hieronder het Venn diagram voor de verzameling.
"vereniging van A en B (bron wikipedia)"



En zo kun je ook het complement van A laten zien:
"complement van A (bron wikipedia)"

4.0.2 Conditionele kans

Een conditionele kans wordt geschreven als $P(A|B)$ en kun je lezen als "Wat is de kans op meting A gegeven dat B is gemeten.". Let op dat $P(A|B) \neq P(B|A)$! Een sprekend voorbeeld hiervan is de volgende. De kans dat een persoon zwanger is gegeven dat de persoon een vrouw is, P(zwanger|vrouw), is niet gelijk aan de kans dat iemand een vrouw



is gegeven dat de persoon zwanger is, P(vrouw|zwanger). De laatste kans is duidelijk gelijk aan 1, als je zwanger bent ben je zeker een vrouw. De eerste kans is een stuk kleiner!

De conditionele kans kunnen we berekenen met:

$P(A|B) = \frac{P(A \cap B)}{P(B)}$.

De noemer in deze vergelijking, $P(B)$, noemen we ook wel een normalisatie term. De kans $P(A \cap B)$ moet genormaliseerd worden naar de kans $P(B)$, immers het is al een gegeven dat B waar is.

Visueel is dit wellicht het meest eenvoudige om te zien. Als het gegeven is dat de uitkomst in het deelgebied B ligt, dan is de kans dat het ook de waarde A bezit gelijk aan het oppervlak van de overlap tussen A en B gedeeld door het oppervlak van B. Immers dat het B is weten we al, dus we moeten alle kansen normaliseren naar B.

4.0.3 Bayes theorema

Met behulp van de conditionele kans formule kunnen we nu Bayes theorema afleiden.
Het combineren van de formules van $P(A|B)$ en $P(B|A)$:

$P(A \cap B) \equiv P(B \cap A) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$

geeft:

$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Dit theorema maakt het mogelijk om nieuwe informatie toe te voegen aan de kennis van de kans. In module 1 hebben we het kort over Bayesiaanse kans definitie gehad. Dit theorema staat centraal in Bayesiaanse kans. Het is wel belangrijk om te weten dat deze wiskundige vergelijking ook opgaat in de Frequentist benadering van kans.

Bekijk ook even het kennisclipje over Extra Kansrekenregels op Canvas!

Voorbeeld van gecorreleerde fouten.

Je meet 1 lengte van een perfecte vierkant en rekent het oppervlakte uit. De formule voor het oppervlakte van een vierkant is A = l1*l2

Hoofdstuk 5

Opdrachten module 2

Tijdens laptopcolleges 3 en 4 werken we aan het de opdrachten van module 2. In deze module gaan we werken aan de volgende opdrachten.

- M2.1 Grote Aantallen II **
- M2.2 Meesjes ****
- M2.3 Halfwaardedikte II ***

De sterren geven een indicatie voor hoeveel werk een opdracht is. Let op dat je deze week goed plant!

De antwoorden van de opdrachten moet je invoeren in dit template en als **pdf** bestand inleveren samen met de code voor de 3 opdrachten 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.

Vergeet niet om ook even te kijken naar de oefen opgaves ter voorbereiding van de tweede tussentoets die aan het einde van het derde hoorcollege plaats vindt.

Veel succes!

5.1 Opdracht M2.1 Grote Aantallen II **

We gaan verder kijken naar de ton met kogels uit opgave M1.4. In dit opgave begonnen we met een ton met 80 kogels en berekenden we het gemiddelde, $g_n = \bar{m}_n$ over de eerste n kogels van de set. Zo kregen we de distributie van g_n versus n, net als in opgave M1.4. Voordat je verder gaat, controleer eerst even in ANS of je dit goed hebt gedaan en corrigeer eventueel je fouten.

We gaan nu naar meerdere tonnen kijken, steeds met 80 kogels en uit dezelfde fabriek. We gaan steeds de waardes van g_n opnieuw berekenen, voor elke ton weer. Als we alle gemiddelde waardes g_n van één ton hebben berekend, dan beginnen we weer opnieuw met

een nieuwe ton met 80 kogels. Dit herhalen we 100 keer.
We gaan er in deze opgave stap voor stap doorheen.

Maak eerst 100 verschillende datasets. Elke dataset is een ton met 80 kogels. Dit kan je doen door steeds een andere seed mee te geven aan de datasetgenerator:

```
datasets = [ds.DataSetGroteAantallen(i) for i in range(0,100)]
```

Je hebt nu een `list` die `datasets` heet met 100 items. Elke item is een dataset met elk 80 meetwaardes.

- **M2.1a) Maak nu eerst een histogram van *alle eerste* elementen, m_1, van de 100 datasets. Zorg dat je histogram er netjes uitziet.**

- **M2.1b) Wat is het gemiddelde, g_1, en de standaarddeviatie s_1 van dit histogram? Denk bij het noteren aan de eenheden en de juiste notatie!**

We gaan nu experimenten vergelijken waarin we steeds het gemiddelde over de eerste 10 metingen (g_{10}) hebben berekend.

- **M2.1c) Bereken voor elk van de 100 datasets het gemiddelde over de eerste 10 metingen en laat de distributie van deze gemiddeldes g_{10} zien in een histogram.**

- **M2.1d) Bereken van deze distributie het gemiddelde $\bar{g_{10}}$, dit is het gemiddelde van de gemiddeldes g_{10}. Bereken ook de standaarddeviatie van de gemiddeldes $s_{g_{10}}$ (de standaarddeviatie van de gemiddeldes g_{10}).**

We gaan dit nu herhalen voor met verschillende groottes van de steekproef n. Maak een functie die de standaarddeviatie s_{g_n} van de 100 berekende gemiddeldes g_n die berekend zijn over de eerste n punten terug geeft.

Roep nu de functie aan voor de volgende waardes van n: 1, 5, 10, 20, 30, 40, 50, 60, 70, 80. Controleer of de punten voor $n = 1$ en $n = 10$ dezelfde resultaten opleveren als dat je net

had.

- **M2.1e) Maak nu een grafiek waarin je de berekende standaarddeviaties s_{g_n} uitzet tegen de grootte van de steekproeven, n.**
- **M2.1f) Maak een nieuwe grafiek waarin je de berekende s_{g_n} uitzet tegen $1/\sqrt{n}$.**
- **M2.1g) Kun je iets zeggen over de grafieken? Beschrijf wat je ziet en probeer daar een conclusie uit te trekken.**

Wat we hebben gedaan in deze opdracht is illustreren wat er gebeurd als we een steeds grotere steekproef nemen.

5.2 M2.2 Meesjes ****

Je vindt helaas een dood meesje in de tuin. Het lijkt op een koolmeesje maar het zou ook een pimpelmeesje kunnen zijn. Deze twee vogeltjes lijken erg veel op elkaar. Er zijn manierenom pimpelmeesjes van koolmeesjes te onderscheiden met behulp van uiterlijke kenmerken. Maar je bent een Natuurkundige en geen Bioloog. Online vind je een dataset met informatie over het massa en de spanwijdte van beide soorten meesjes.

Voordat we aan deze opdracht beginnen moeten we eerst een nieuwe versie downloaden van de DAS_DatasetGenerator.py. Zonder de nieuwe versie werkt deze opgave niet. Download ook het bestand M2.2_Meesjes.py en zorg dat deze in dezelfde folder staat als het `DAS_DatasetGenerator.py` bestand.

We genereren eerst een twee datasets met behulp van de volgende regel code:

```
m_km, span_km, m_pm, span_pm = ds.datasetVogeltjes()
```

De variabelen hebben de volgende betekenis:

```
m_km    : de massa van een koolmeesje in gram
span_km : de spanwijdte van een koolmeesje in cm
```

De laatste twee variabelen zijn de datapunten voor pimpelmeesjes. De twee variabelen van de koolmeesjes horen bij elkaar. Van elk meesje in de dataset zijn zowel de massa als de spanwijdte gemeten. De dataset is zo geordend dat als je het n-de punt uit de `m_km`-lijst bij het n-de punt uit de `span_km`-lijst hoort. Dit zijn de gegevens van het n-de meesje. Pas dus op dat je de lijsten in de juiste volgorde houdt! Voor de twee variabelen van de pimpelmeesjes geldt precies hetzelfde.

We gaan eerst naar de twee massaverdelingen van de meesjes kijken.

- **M2.2a) Plot de massaverdelingen van beide meesjes in een histogram. Laat in een legenda zien welke meesje bij welke kleur hoort. Maak ook een apart histogram waarin je spanwijdtes van de twee soorten meesjes plot. Maak de twee histogrammen netjes af en zorg dat duidelijk is welke distributie bij welk soort meesje hoort.**

  TIP: Gebruik de plot optie `alpha=0.8` zodat je histogrammen wat doorzichtig worden. Zo kan je het achterste histogram ook nog altijd goed zien.

- **M2.2b) Maak een tabel waarin je voor beide soorten meesjes de gemiddeldes, de standaarddeviaties en de varianties noteert. Let goed op de notatie en denk ook even aan de eenheden.**

We meten nu de massa op van het meesje dat je gevonden hebt. Gebruik de volgende regel code om dat te doen:

```
mees_m_laag, mees_m_hoog = ds.meetMassaMeesje()
```

Je krijgt nu een onderwaarde `mees_m_laag` en een bovenwaarde `mees_m_hoog` terug. Deze geven de onzekerheid op de meting aan. Het gemiddelde van deze twee is de gemeten massa, de centrale waarde. De waarde van de massa van de mees ligt **zeker** tussen de boven- en onderwaarde in. NB. Als je een foutmelding krijgt dat `meetMassaMeesje()` niet bestaat controleer dan of je wel een nieuwe `DAS_DatasetGenerator.py` hebt downgeload voor Module 2.

Met deze informatie kunnen we nu met de Frequentist Methode de kans uitrekenen dat onze mees een Koolmeesje is.

- **M2.2c) Gebruik de dataset `m_km` om de kans uit te rekenen dat je een koolmeesje vindt die een massa heeft die in het gebied `mees_m_laag` en `mees_m_hoog` in ligt. Dit noem je ook wel de voorwaardelijke kans $P($**

De frequentist methode, zoals we die hierboven gebruiken, is uiteindelijk een ratio tussen twee getallen. Deze twee getallen hebben een onzekerheid volgens de Poisson verdeling.

- **M2.2e) Schrijf de formule uit hoe de onzekerheden van de noemen en deler zich propageren naar de onzekerheid op de uitgerekende kans. Noteer deze formule en bereken met behulp van deze formule de onzekerheden uit op de kansen die je in M2.2c) hebt berekend.**

Je besluit ook de spanwijdte van de mees op te meten. Misschien geeft dat wel meer uitsluitsel.

```
mees_span_laag, mees_span_hoog = ds.meetLengteMeesje()
```

De output volgt dezelfde logica als hiervoor.

- **M2.2f) Gebruik dezelfde methode als hiervoor om beide kansen $P(w_{\mathbf{obs}}|\mathbf{koolmees})$ en $P(w_{\mathbf{obs}}|\mathbf{pimpelmees})$ uit te rekenen maar nu door (alleen) gebruik te maken van de informatie van de spanwijdtes. Noteer ook de onzekerheden op de uitgerekende kansen.**

- **M2.2g) Op basis van deze informatie, wat denk je nu dat het voor vogeltje is?**

We kunnen nu natuurlijk ook de gecombineerde informatie gebruiken. Hiervoor gaan we eerst de data visualiseren.

- **M2.2h) Maak een tweedimensionale scatterplot die de tweedimensionale dataset van de massa versus de spanwijdte voor zowel de pimpelmezen als de koolmezen.**
  TIP gebruik de opties `'o',markersize=3, alpha=0.4` in de plot functie. Zorg dat beide datasets weer hun eigen kleur hebben en vergeet de legenda niet.

Het valt misschien op dat er een verband lijkt te zijn tussen beide variabelen. We gaan daar eerst naar kijken naar de covariantie en de correlatie tussen de massa en de spanwijdte voor beide vogelsoorten.

- **M2.2i) Bereken de covariantie en de correlatie tussen de massa en de spanwijdte voor zowel de koolmeesje als de pimpelmeesjes meetgegevens.**

- **M2.2j) Als je naar de berekende correlaties kijkt wat valt dan op, wat voor verband zit er tussen de twee variabelen? Als je toch even als een Bioloog nadenkt, is dit dan wat je verwacht?**

We gaan terug naar de kansberekeningen.

- **M2.2k) Combineer nu de gegevens en bereken de kansen $P(m_{\mathbf{obs}}$ en $w_{\mathbf{obs}}|$koolmees$)$ en $P(m_{\mathbf{obs}}$ en $w_{\mathbf{obs}}|$pimpelmees$)$.**

- **M2.2l) Welk vogeltje denk je nu dat het is? Beredeneer je antwoord.**

Na al deze berekeningen lopen we een eindje in de tuin. Op de plek waar we eerder het meesje aantroffen zit nu een ander meesje hartstochtelijk te zingen. Aan de zang hoor je direct dat dit een pimpelmeesje is. Je schat in dat er een kans is van 90% dat dit pimpelmeesje bij het andere meesje hoorde, en dat dat dus ook een pimpelmees is.

- **M2.2m) Bereken nu de kans dat het inderdaad een pimpelmeesje is geweest: $P($pimpelmees$|m_{\mathbf{obs}}$ en $w_{\mathbf{obs}})$. Bereken hier alleen de centrale waarde.**
  TIP: Maak hierbij gebruik van de vergelijking van Bayes. Om $P(m_{\text{obs}}$ en $w_{\text{obs}})$ te berekenen kun je gebruiken maken van de volgende formule: $P(C) = P(C|D) \cdot P(D) + P(C|\text{niet } D) \cdot P(\text{niet } D)$.

5.3 M2.3 Halfwaardedikte II ***

We gaan nu terug naar het experiment uit opgave M1.5 waarbij we de halfwaardedikte van lood onderzoeken bij een bepaalde gamma-bron. We gaan de onzekerheid op het meetresultaat, d_{half} onderzoeken.

In opgave M1.5 gebruikten we een methode om de halfwaardedikte te bepalen waarbij we steeds de ratio tussen het aantal counts zonder lood N_0 en een waarde met lood N_{half} uitrekende. Zodra deze ratio onder de 0.5 komt nemen we x als de halfwaardedikte.

- **M2.3a) Wat is de onzekerheid op de ratio R? Bereken deze door gebruik te maken van de onzekerheden op N_0 en N_{half} en de regels voor propagatie van ongecorreleerde fouten (Deze kan je hier vinden). Schrijf je berekening helemaal uit.**

Een **schatter** is een recept om de waarde van een parameter af te schatten. De parameter die we hier willen bepalen is de halwaardedikte van lood (voor de energie van onze bron).

De schatter is in dit geval d_{half}.

Nu gaan we het experiment 50 keer herhalen en gaan we kijken naar de distributie van de gevonden halfwaardediktes. We gaan uit deze distributie de standaarddeviatie halen en dit gebruiken om de onzekerheid op de gevonden dikte d_{half} te bepalen.

Schrijf een loop waarin je 50x een nieuwe dataset genereert waaruit je 50x opnieuw een halfwaardedikte bepaald. Om 50 unieke dataset te maken moet je steeds de zogeheten ***seed*** veranderen. Dat kan je doen door een seed mee te geven aan de DAS dataset generator:

```
for j in range(0,50) :                                          1
    counts, diktes = ds.DataSetHalfwaardeDikte(j)               2
```

Binnen deze loop maak je 50 unieke datasets aan waarbij de counts die gemeten worden steeds worden gevarieerd volgens de Poisson statistiek.

- **M2.3b) Maak een histogram waarin je de gevonden halfwaardediktes van de 50 verschillende experimenten laat zien. Zorg dat het histogram de distributie netjes laat zien en dat de as-labels goed zijn aangemaakt.**
  TIP: de binning in het histogram luistert nauw doordat er alleen bepaalde uitkomsten van de halfwaardedikte mogelijk zijn. Reken precies uit wat de range en de binning moet zijn in het histogram om te voorkomen dat je lege bins midden in de distributie krijgt.

- **M2.3c) Ziet de distributie eruit zoals je verwacht had? Beredeneer je antwoord.**

- **M2.3d) Bepaal nu het gemiddelde van de meetuitkomsten en de standaarddeviatie van de distributie.**

- **M2.3e) Zeggen deze getallen ook iets of de gemeten waardes gemiddeld te hoog of te laag uitkomen. Beredeneer je antwoord.**

We gaan nu kijken hoe zuiver de meting is. De onzuiverheid is gedefinieerd als het verschil tussen de echte waarde en de gemiddelde gemeten waarde. Bij gesimuleerde data kunnen we dit onderzoeken, daarvan kunnen we het meetresultaat vergelijken met de initiële waardes die we hebben gebruikt in de simulatie.

Om de zuiverheid van ons experiment te bepalen gaan we dus de bepaalde halfwaardedikte te vergelijken met de initiële halfwaardedikte die gebruikt is om de data te simuleren. Roep hiervoor de volgende functie aan in de dataset generator:

```
metingen, diktes, d_true = ds.DataSetHalfwaardeDikteVariatie(s,d_input)
```

Je geeft twee variabelen mee aan de functie: de seed (`s`) en een waarde(`d_input`). We komen er zo op terug wat deze variabelen betekenen. De functie geeft drie objecten terug. De eerste twee zijn de lists met de counts en de looddikte (zoals je eerder ook terugkreeg), de derde variabele is halfwaardedikte die gebruikt is als input voor de simulatie. Dit noemen we meestal de *true* waarde in simulaties vandaar dat we hem `d_true` noemen. Met de variabele `d_input` kunnen we nu de input waarde van de simulatie controleren. In principe is `d_input` gelijk aan `d_true`, tenzij je de waarde -1 kiest.

Met deze dataset generator gaan we nu de zuiverheid van onze meting bestuderen.

- Kijk eerste eens naar wat de *true* waarde was in je datasets die je hierboven hebt gebruikt! Als je voor `d_input` nu -1 invult krijg je de halfwaardedikte die gebruikt is voor het genereren van de 50 datasets die je eerder in deze opdracht hebt gebruikt.

- **M2.3f) Hoe groot is de onzuiverheid van ons experiment? Vergelijk hiervoor de gemiddelde bepaalde halfwaardediktes van de 50 experimenten met de `d_true`.**

Nu kun je het gedrag bekijken over meerdere waardes rond de `d_true` waarde. Plaats een grote loop over je hele code en varieer de `d_input` waarde bijvoorbeeld met 5 of 10 procent rond je aanvankelijke waarde. Voor elke setting van `d_input` bepaal je over 50 experimenten het gemiddelde van de bepaalde waardes van d_{half}.

- **M2.3g) Zet de gevonden gemiddelde waardes zet je in een grafiek uit tegen de gekozen waardes van`d_input`. Let goed op de leesbaarheid van je grafiek en zorg dat je makkelijk kunt aflezen waar de zuivere meting zou liggen (dus als d_half = d_input).**

- **M2.3h) Is de onzuiverheid altijd constant of varieert die afhankelijk van de halfwaardedikte?**

- **M2.3i) In dit geval simuleren we het experiment. Zou je een methode kunnen bedenken om de onzuiverheid van je experiment te onderzoeken bij een echte meting?**