Data Analyse en Statistiek

Hella Snoek & Marthe Schut

Inhoud

0.1 Introductie Module 1

Statistische data analyse is een belangrijk onderdeel in vele werkvelden. Als student en later wellicht ook als wetenschapper zul je te maken krijgen met het verzamelen en interpreteren van data bij het het practikum, bij het doen van onderzoek, of juist bij het begrijpen van de interpretatie van andermans resultaten.

- Wanneer kun je zeggen dat een hypothese moet worden verworpen of bewijs je juist dat deze correct is?
- Hoe moet je inschatten of je meetnauwkeurigheid goed genoeg is? Wanneer heb je eigenlijk genoeg data verzameld?
- Hoe kun je een zo experiment ontwerpen dat je een hypothese kunt onderzoeken.
- Hoe kom je erachter wat jouw hypothese toetsbaar maakt - in welke observabele onderscheidt zij zich voldoende van andere hypotheses?

Alle kennis die we tot nu toe hebben over de Natuur- en Sterrenkunde is tot stand gekomen met het uitvoeren van experimenten en het analyseren van de uitkomsten hiervan. Voor het bestuderen van Natuurkundige en Sterrenkundige theorieën is niet persé kennis nodig van de statistiek en van data analyse technieken. Voor het uitvoeren van wetenschap, het vinden van bewijzen voor nieuwe theorieën is kennis hiervan echter essentieel.

Bij het presenteren van onderzoeksresultaten is het belangrijk om helder uit te kunnen leggen hoe het onderzoek precies is uitgevoerd, hoe de metingen zijn verkregen en wat de resultaten zijn. Vaak maken we hierbij gebruik van histogrammen, grafieken en tabellen. Om een hypothese te toetsen moeten we metingen ook kunnen interpreteren. Hiervoor zijn verschillende methodes, bijvoorbeeld kunnen we de data proberen te 'fitten' met een functie, een wiskundige vergelijking. Bij al deze methodes speelt statistiek een belangrijke rol.

In deze cursus zullen we vaardigheden gaan leren voor data analyse en statistiek.

Deze week beginnen we met een aantal basisbegrippen in de beschrijvende statistiek. We gaan kijken naar het gemiddelde, variantie, de standaardafwijking, en coëfficiënt van variantie. We leren over hoe we meetresultaten moeten presenteren, het gebruik van de wetenschappelijke notatie en hoe we ze kunnen visualiseren. We gaan in op het begrip meetonzekerheid.
Ook maken we een begin met kansrekening en kansdichtheidsverdelingen.

Niet elk van deze onderwerpen is even moeilijk. Let goed op dat je genoeg tijd overhoudt om de introductie van de kanstheorie te bestuderen.

We werken in de werkcolleges aan de opdrachten van deze module M1. Je vindt in het schema wanneer je deze moet inleveren. Vergeet ook niet te kijken naar het oefenmateriaal voor de eerste verplichte tussentoets die volgt aan het einde van het tweede hoorcollege.

0.2 Basisbegrippen in de statistiek

1.

0.2.1 Datasets beschrijven

Als we een set metingen (data) hebben verzameld kunnen we deze op verschillende manier gebruiken. Vaak willen we bepaalde kenmerken van de dataset weten. Stel we hebben een dataset met de temperatuur op elk van de 37 meetpunten van het KNMI in Nederland in de afgelopen twintig jaar. Het is dan niet zo inzichtelijk om dit aan medewetenschappers te presenteren d.m.v. een enorme tabel (elke 10 minuten wordt een meting gedaan door de weerstations) met de mededeling 'dit was de temperatuur in de afgelopen twintig jaar'. Uit deze dataset kun je natuurlijk een enorme hoeveelheid informatie halen. Bijvoorbeeld wat is de koudste temperatuur die in de afgelopen 20 jaar in Nederland is gemeten. Maar ook: Wat is de gemiddelde temperatuur in de maand Juli. Of: Hoeveel kouder zijn de winters in het binnenland ten opzichte van de kust regio's.

In de secties hieronder behandelen we verschillende veelvoorkomende definities van kenmerken van data.

0.2.1.1 Populatie en steekproef

Voordat we het gaan hebben over de kenmerken van data is het belangrijk om te kijken naar de data zelf. Waar komt die vandaan? We maken hierbij onderscheid tussen de **populatie** en een **steekproef**.

Een **populatie** bestaat uit alle personen/dieren/objecten binnen de groep waarin we geïnteresseerd zijn. Dit zouden bijvoorbeeld *alle* mensen in Nederland kunnen zijn tussen de 30 en 40 jaar, of *alle* lieveheersbeestjes die in Noorwegen leven. Nu is het zo dat het vaak lastig is om van *alle* personen/dieren/objecten (hierna uniform aangeduid met 'elementen') van een groep gegevens te verzamelen. Het kost bijvoorbeeld erg veel tijd (en geld) om data te verzamelen over alle personen tussen de 30 en 40 jaar in Nederland (of om alle lieveheersbeestjes in Noorwegen te vangen). Het is dan veel makkelijker om data over een deel van deze groep te verzamelen en om zo toch iets te kunnen zeggen over de gehele doelgroep. Zo zouden we bijvoorbeeld data kunnen verzamelen van een willekeurige selectie van 200 personen in Nederland tussen de 30 en 40 jaar. Dit wordt een *steekproef* genoemd, de deelgroep wordt in het Engels vaak aangeduid met een *sample*. Een steekproef is dus een gedeelte van de populatie. Vaak is het trouwens zelfs helemaal niet mogelijk om de hele populatie te meten. Denk bijvoorbeeld maar eens aan de gemiddelde massa van een ster. Dan zouden we deze meting moeten verrichten voor alle sterren in het universum.

We maken onderscheid in de namen en de notatie van de kenmerken van data. Kenmerken van meetgegevens (data) van een populatie noemen we **parameters**, kenmerken van steekproeven noemen we **statistieken**. Het is belangrijk om onderscheid te maken. Als we bijvoorbeeld de gemiddelde leeftijd willen weten van alle eerstejaars Natuur- en Sterrenkunde studenten in Amsterdam dan maakt het uit of we de gegevens hebben verzameld van alle eerstejaars of dat we de gemiddelde leeftijd inschatten door de gegevens te noteren van de studenten uit je eigen werkgroep. In het eerste geval hebben we gegevens van de hele populatie en spreken we van een parameter en weten we de uitkomst exact. In het tweede geval hebben we een steekproef gedaan van een selectie van de eerstejaars, we spreken dan van een statistiek en op deze statistiek komt een onzekerheid. We hebben immers niet alle

informatie van de populatie en het kan zijn dat het gemiddelde van de steekproef afwijkt van het gemiddelde van de gehele populatie. Het is dus belangrijk om je te realiseren of je de gegevens bekijkt van een steekproef of een populatie als je de resultaten interpreteert.

Als je een steekproef neemt is het belangrijk om op twee dingen goed te letten: de grootte van de steekproef en hoe representatief deze is. Je kunt je voorstellen dat als we de lengte van drie mensen in Nederland meten, we nog niet zoveel kunnen zeggen over de lengte van de gehele populatie die bestaat uit alle mensen in Nederland. Als we de lengte van 1000 mensen zouden meten dan krijgen we al een beter beeld van de verdeling van lichaamslengte in Nederland, en kiezen we 100000 mensen dan krijgen we een nog veel beter beeld van de verdeling. Hoe groter de steekproef, hoe nauwkeuriger de statistiek is die we willen weten. (We zeggen dan vaak dat we *meer statistiek* hebben.)

Ook is het belangrijk hoe we de steekproef nemen. Als we bijvoorbeeld de lengte gegevens van 1000 mensen nemen dan krijgen we een vertekend beeld als we hiervoor de leden van de Nederlandse Basketbal vereniging uitnodigen, of de gegevens van 1000 kleuters hiervoor gebruiken. Je moet dus altijd goed kijken of de steekproef de je neemt wel representatief is voor de hele groep.

0.2.2 Veel gebruikte parameters en statistieken

0.2.2.1 Het gemiddelde

Het gemiddelde van een dataset geeft een maat voor het centrum van de waarden die de dataset aanneemt. We onderscheiden het populatiegemiddelde (parameter) en het steekproefgemiddelde (statistiek). Hoe groter de steekproef hoe meer het gemiddelde van de steekproef overeenkomt met het populatiegemiddelde.

Het gemiddelde kun je berekenen door alle waardes in de dataset te sommeren en te delen door de grootte van de dataset. We maken onderscheid in de notatie voor het gemiddelde van een steekproef en die van het populatiegemiddelde.

Het steekproef gemiddelde $\bar{x}$ (x-streep of in het Engels: x-bar) van een dataset is de som van de waarden $x_1, \ldots, x_n$ in de set gedeeld door het aantal datapunten in de steekproef: n:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Het steekproef gemiddelde wordt zo vaak gebruikt dat dit veelal wordt aangeduid als 'het gemiddelde'. Voor het gemiddelde wordt ook vaak de 'vishaak-notatie' gebruikt: $< x >$.

Het populatiegemiddelde wordt als volgt genoteerd:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

Hierbij is N het aantal elementen in de populatie, en zijn $x_i, \ldots, x_N$ de waardes van de grootheid in de populatie. Let op dat voor de steekproefgrootte n wordt gebruikt en voor de populatiegrootte N. Een andere veel gebruikte notatie voor het populatiegemiddelde is $E(x)$ waar de E van het Engelse woord *expectation* komt. Ook kun je een subscript toevoegen om aan te geven van welke grootheid je het gemiddelde berekent, bijvoorbeeld hier μ_x.

Je ziet dat het steekproef gemiddelde erg lijkt op de uitdrukking voor het populatiegemiddelde. Het verschil is dat het steekproefgemiddelde niet persé gelijk is aan de verwachtingswaarde van de populatie. Het is wel zo dat, hoe beter de steekproef overeenkomt met de populatie, des te dichter komt het steekproef gemiddelde bij de verwachtingswaarde van de populatie. Met behulp van een goed uitgevoerde steekproef kan het statistische gedrag van een populatie dus benaderd worden.

Voorbeeld:

Stel je voor dat we de volgende steekproef hebben: $X = \{-5, 1, 14, 12, 0\}$. De gemiddelde waarde voor de data is nu dus
$\bar{x} = \frac{1}{5} \cdot (-5 + 1 + 14 + 12 + 0) = \frac{1}{5} \cdot 22 = 4.4$

0.2.2.2 De mediaan

De mediaan is een maat voor het midden van de elementen in een gesorteerde dataset of verdeling. De mediaan is zo gedefinieerd dat je precies 50% kans hebt om een waarde te vinden die lager is dan de mediaan en 50% kans om een waarde te vinden die hoger is dan de mediaan.

Als we alle datapunten in een dataset sorteren van lage naar hoge waarde, dan is de mediaan de waarde van het element in het midden van de set. Is er sprake van een even aantal elementen dan is de mediaan de gemiddelde waarde van de twee elementen in het midden van de set.

Voorbeeld:

Stel dat we de volgende dataset hebben: $X = \{13, 11, 10, 14, 12, 9\}$.
Het eerste wat we moeten doen om de mediaan te vinden is de dataset sorteren: $\{9, 10, 11, 12, 13, 14\}$.
We hebben een dataset met een even aantal datapunten, de mediaan ligt hier dus tussen twee waardes in: $\frac{(11+12)}{2} = 11.5$.

De mediaan en het gemiddelde *kunnen* dezelfde waarde hebben, maar dat hoeft niet zo te zijn. Voor het voorbeeld hierboven is dat wel het geval (reken maar na). Maar voor de

dataset uit het voorbeeld voor het berekenen van het gemiddelde is dit niet zo. Kijk maar!

Voorbeeld:

We bekijken de steekproef $X = \{-5, 1, 14, 12, 0\}$. Het gemiddelde was berekend op 4.4. We gaan nu kijken waar de mediaan ligt. Eerst sorteren we de dataset: $\{-5, 0, 1, 12, 14\}$. Dit is een oneven dataset en de mediaan ligt dus op de middelste waarde van de gesorteerde dataset: 1.

Voor symmetrische datasets zijn het gemiddelde en de mediaan altijd gelijk aan elkaar, voor asymmetrische datasets is dit niet het geval. Bij een symmetrische dataset is de data precies gespiegeld rond het gemiddelde. Dit is makkelijker uit te leggen aan de hand van datadistributies. We komen hier later op terug.

0.2.2.3 De modus

De modus van een dataset is de waarde die met de grootste frequentie in de dataset voorkomt. Hebben we bijvoorbeeld de dataset

$$2, 2, 3, 4, 7, 7, 7, 9 \tag{3}$$

dan komen de 3, de 4 en de 9 elk één keer voor, het getal 2 komt twee keer voor en het getal 7 komt drie keer voor. Het meest voorkomende getal is dus de 7 en dit is de modus van de dataset. Als een dataset één modus heeft dan wordt deze *unimodaal* genoemd.

Het komt ook voor dat er twee of meer getallen zijn die vaker voorkomen dan andere waardes. Een dataset met twee getallen als modus wordt ook wel *bimodaal* genoemd, een dataset met meer dan twee getallen als modus wordt *multimodaal* genoemd.

Een voorbeeld van een bimodale dataset is:

$$1, 2, 3, 3, 4, 4, 4, 5, 6, 11, 11, 11, 15 \tag{4}$$

zowel het getal 4 als het getal 11 komen drie keer voor in de set. De set is dus bimodaal met modus 4 en modus 11.

Bij sommige soorten dataverdelingen is het gebruikelijker om over de modus te praten dan over het gemiddelde of de mediaan. Een voorbeeld hiervan is de Landau distributie die een slecht gedefinieerd gemiddelde of mediaan kent door een lange staart in de distributie.

Voor unimodale symmetrische distributies ligt het gemiddelde, de mediaan en de modus precies op dezelfde plek.

0.2.3 Spreiding van data

De spreiding geeft een beeld van de mate waarin datapunten in een set verspreid zijn. Er zijn verschillende maten om de spreiding van een dataset mee aan te geven. Hieronder zullen we **de spreidingsbreedte** (ook wel de *range*), **de variantie**, **coëfficiënt van variantie** en **de standaarddeviatie** (ook wel de *standaardafwijking*) bespreken.

0.2.3.1 Spreidingsbreedte (range)

De range is de afstand tussen de hoogste en de laagste waarde in een dataset. Hebben we bijvoorbeeld de dataset

$$50, 70, 72, 76, 76, 80, 120 \tag{5}$$

dan is de range van deze dataset gelijk aan $120 - 50 = 70$.

De range geeft dus aan hoe breed de dataset in totaliteit is. De range is niet altijd een handige maat voor de spreiding van een dataset. Zo zouden we bijvoorbeeld de volgende dataset kunnen hebben:

$$1, 2, 3, 4, 5, 5, 5, 6, 6, 6, 7, 7, 10 \tag{6}$$

De range is in dit geval $10 - 1 = 9$. Maar stel dat we een foutieve meting doen (of we maken een typefout in het overnemen van de data), en we hebben de volgende dataset:

$$1, 2, 3, 4, 5, 5, 5, 6, 6, 6, 7, 7, 10, 30 \tag{7}$$

De range wordt nu $30 - 1 = 29$. Dus onder invloed van één foutief datapunt geeft de range nu een veel grotere mate van spreiding aan.

0.2.3.2 Standaarddeviatie en variantie

De standaarddeviatie (ook wel de standaardafwijking) geeft aan in welke mate de data verspreid is rondom het gemiddelde van de dataset. Dit geeft met name ook een maat voor de spreiding van de datapunten onderling. Hoe groter de standaarddeviatie des te groter is de spreiding tussen de afzonderlijke punten. De standaarddeviatie voor de populatie wordt aangeduid met σ, voor een steekproef noteren we dit met s.

De variantie, *var*, is direct gerelateerd aan de standaarddeviatie, namelijk de variantie is gelijk aan de standaarddeviatie in het kwadraat. Voor de populatie geldt dus var $= \sigma^2$. De variantie van een steekproef noteren we met s^2.

De variantie en standaarddeviatie van een populatie kunnen worden berekend met de volgende formule:

$$var = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{8}$$

of in het geval van de steekproef:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{9}$$

Let op dat de eenheid van de variantie het kwadraat is van de eenheid van x. In het geval dat je bijvoorbeeld lengtes van luciferstokjes hebt opgemeten, dan zullen de waardes in cm zijn genoteerd. De variantie heeft dan de eenheid cm^2. Dat kan soms best onhandig zijn, vandaar dat we vaker de standaarddeviatie gebruiken. De standaarddeviatie heeft altijd dezelfde eenheid als de originele elementen van de dataset.

0.2.3.3 Variatiecoëfficiënt

De variatiecoëfficiënt wordt ook wel de relatieve standaardafwijking genoemd. De coëfficiënt van variatie geeft, net zoals de standaardafwijking en de variantie, een maat voor de spreiding van de populatie of dataset.

De variatiecoëfficiënt wordt gegeven door de verhouding tussen de standaardafwijking en het gemiddelde. Voor een populatie is de coëfficiënt van variantie c_v dan:

$$c_v = \frac{\sigma}{\mu} \tag{10}$$

Met σ de standaardafwijking van de populatie en μ het populatiegemiddelde.

De steekproef variantie $\hat{c_v}$ wordt gegeven door:

$$\hat{c_v} = \frac{s}{\bar{x}} \tag{11}$$

Met s de standaardafwijking van de steekproef en $\bar{x}$ het steekproef gemiddelde.

Het verschil met de variantie en de standaardafwijking is dat de variatiecoëfficiënt dimensieloos is. Dit is bijvoorbeeld handig als er meerdere datasets vergeleken moeten worden die verschillende eenheden hebben. Ook als de gemiddelde waarden van verschillende datasets erg uiteen liggen is het beter om de variatiecoëfficiënt te gebruiken i.p.v. de standaardafwijking.

Een nadeel van het gebruik van de variatiecoëfficiënt is dat er gedeeld wordt door het gemiddelde. Als dit gemiddelde een heel kleine waarde heeft, dicht bij nul, dan is de variatiecoëfficiënt slecht gedefinieerd.

0.2.4 Samenvatting

kenmerk	populatie (*parameter*)	steekproef (*statistiek*)
grootte	N	n
gemiddelde	$\mu = \frac{1}{N}\sum_i^N x_i$	$\bar{x} = \frac{1}{n}\sum_i^n x_i$
standaarddeviatie	$\sigma = \sqrt{\frac{1}{N}\sum_i^N (x_i - \mu)^2}$	$s = \sqrt{\frac{1}{n}\sum_i^n (x_i - \bar{x})^2}$
variantie	$var = \sigma^2$	s^2
variatiecoëfficiënt	$c_v = \sigma/\mu$	$\hat{c_v} = s/\bar{x}$

0.2.5 Voorbeelden

We berekenen de eigenschappen van een aantal datasets als voorbeeld.

Voorbeeld:

We hebben de volgende dataset van een populatie:
$Y = \{285, -20, 31, 60, 12, 53, 133\}$.

De grootte is dus N = 7.
Om de mediaan te bepalen sorteren we eerst de datapunten van klein naar groot:
-20,12,31,53,60,133,285
Het is een even aantal datapunten en de mediaan ligt tussen 53 en 60 in. Dit komt dan uit op 56.5.
De spreidingsbreedte is $285 - -20 = 305$.
Het gemiddelde $\mu_Y = \frac{1}{7} \cdot (285 - 20 + 31 + 60 + 12 + 53 + 133) = 79.1$
De standaarddeviatie is: $\sigma_Y^2 = \frac{1}{7} \cdot [(285 - 79.1)^2 + (-20 - 79.1)^2 + (31 - 79.1)^2 + (60 - 79.1)^2 + (12 - 79.1)^2 + (53 - 79.1)^2 + (133 - 79.1)^2] = 8997.6$ geeft $\sigma_Y = 94.9$.
De variantie $var_Y = 8997.6$. De variatiecoëfficiënt $c_v = 1.20$.

Het tweede voorbeeld gaat over een steekproef:

Voorbeeld:

Stel we hebben een steekproef gedaan van de lengte van eerstejaars studenten. De volgende dataset is hiervoor verzameld:
$L = \{1.90\text{ m}; 1.72\text{ m}; 1.61\text{ m}; 1.84\text{ m}; 1.79\text{ m}\}$.
De grootte van de steekproef: $n = 5$.
De spreidingsbreedte is 1.90m-1.61 m = 39 cm.
De mediaan ligt in het midden van de gesorteerde dataset. Dit is 1.79 m.
Het gemiddelde $\bar{L} = 1.77$ m.
De variantie is:
$s^2 = \frac{1}{5} \cdot [(1.90 - 1.77)^2 + (1.72 - 1.77)^2 + (1.61 - 1.77)^2 + (1.84 - 1.77)^2 + (1.79 - 1.77)^2] = 0.0100\text{m}^2$

De standaarddeviatie is $s = 0.10$ m.
De variatiecoëfficiënt is $\hat{c_v} = 0.0057$

0.3 Het correct noteren van resultaten

1.

Voordat we verder gaan is het belangrijk om even in te gaan in het onderwerp significantie en de wetenschappelijke notatie. Dit gaat over hoe noteren we een resultaat. Het is goed om even hierbij stil te staan.

Stel dat we een lang meetlint hebben met een millimeter verdeling. We meten de lengte van een lange plank op. We noteren 253.3 cm. We hebben de plank goed kunnen opmeten en er staat een millimeter verdeling op het meetlint. We hebben de meting tot op de millimeter nauwkeurig gedaan. Stel nu dat we met hetzelfde meetlint de hoogte van een struik opmeten. Is het dan oké voor deze hoogte ook de millimeters te noteren? Het opmeten zal waarschijnlijk wel lastig worden. Waar begint bijvoorbeeld de stam van het struikje. De aarde zal wel niet helemaal glad zijn. En lukt het wel om loodrecht op de aarde te meten?

Hoeveel getallen we noteren zegt vaak iets over nauwkeurig we denken het resultaat te weten. Meer hierover komt later terug in het stukje over meetonzekerheid.

Een ander voorbeeld is als we het gemiddelde van drie stokken willen uitrekenen. De stokken zijn 45, 50 en 54 cm lang. We rekenen het gemiddelde uit met onze rekenmachine en we kopiëren het resultaat: 49.66666666 cm. Het lijkt nu of we het resultaat super-nauwkeurig weten terwijl we voor de stoklengtes alleen de centimeters hebben genoteerd. Dat klopt natuurlijk niet!

Voor het noteren van wetenschappelijke resultaten maken we nu hier afspraken die we voor de bachelor vakken gebruiken. Hetzelfde geldt voor het visualiseren van data, daarvoor maken we in het volgende hoofdstuk afspraken. Het is goed om je te realiseren dat er soms wat kleine verschillen kunnen zijn in de afspraken omtrent de visualisatie en de notatie. Als je later in je bachelor een project gaat doen kan het zijn dat de consensus over het presenteren van resultaten net iets anders ligt. Voor nu spreken we de regels af zoals die hieronder volgen.

We beginnen met het uitleggen van wat begrippen die we nodig hebben om de afspraken uit te kunnen leggen.

0.3.0.1 Significantie en precisie

Meetwaardes moeten met de juiste **significantie** worden genoteerd. De *significantie* is de nauwkeurigheid waarmee een getal/waarde wordt weergegeven. Vaak wordt gedacht dat het aantal decimale cijfers de nauwkeurigheid aangeeft, maar dit is technisch gezien de

precisie waarmee de (meet)waarde wordt aangegeven. De nauwkeurigheid (significantie) van een getal zegt welke cijfers in het getal er iets toe doen. Cijfers zonder betekenis tellen we niet mee bij de significantie.

Om de significantie en de precisie te bepalen is het belangrijk om op de nullen te letten en de positie van de punt.

Voor de **significantie** geldt:

- Nullen aan de linkerkant doen niet mee. Het getal 0.0056 heeft bijvoorbeeld twee significante cijfers.
- Nullen aan de linkerkant voorafgegaan door een getal doen wel mee met de significantie. Het getal 100.004 heeft zes significante cijfers.
- Nullen aan de rechterkant doen wel mee met de significantie. Zo heeft 10.34000 zeven significante cijfers.
- Een uitzondering op de tweede regel zijn getallen zoals 300, 4000, 570 etc. Deze getallen zijn weergeven zonder decimalen waardoor het onduidelijk is of daadwerkelijk de waarde van 300 respectievelijk 4000 en 570 is gemeten, of dat dit met een hogere of juist lagere precisie is gebeurd. De afspraak is dat als een getal op deze manier wordt weergegeven met nullen rechts, deze nullen niet meedoen met de nauwkeurigheid. De getallen 300 en 4000 hebben bijvoorbeeld allebei een significantie van 1. Het getal 570 heeft twee significante cijfers. Om deze getallen met een ander aantal significante cijfers weer te geven wordt vaak de *wetenschappelijke notatie* gebruikt. Hier komen we later op terug.

De **precisie** van een getal wordt gegeven door het aantal cijfers achter de punt.

Voorbeeld:

Een aantal voorbeelden

- Het getal 7.134 heeft in totaal 4 significante cijfers, de precisie is 3.
- Het getal 0.576 heeft 3 significante cijfers, de precisie is ook 3.
- 0.001 heeft 1 significant cijfer, de precisie is 3.
- 1.001 heeft 4 significante cijfers, de precisie is 3.
- 2.4500 heeft 5 significante cijfers, de precisie is 4.

In het voorbeeld hierboven zie je dat de getallen (bijna) allemaal dezelfde precisie hebben, maar wel een variatie aan significante cijfers.

0.3.0.2 Wetenschappelijke notatie

Een veel gebruikte manier om getallen en meetresultaten weer te geven is met behulp van de wetenschappelijke notatie. Bij de wetenschappelijke notatie wordt elk getal in de vorm $A \times 10^n$ opgeschreven. Een voordeel van deze notatie is dat je hiermee ook hele kleine getallen en hele grote getallen op een makkelijke manier op kunt schrijven. We geven een voorbeeld:

Voorbeeld:

Voorbeeld klein getal We willen het getal 0.000000000004563 opschrijven met twee significante cijfers. Nu kunnen we natuurlijk 0.0000000000046 opschrijven maar als we dat vaak moeten doen kost dat veel ruimte (en werk). In de wetenschappelijke notatie ziet dit getal met twee significante cijfers er als volgt uit:
$0.000000000004563 = 4.6 \cdot 10^{-12}$

In het voorbeeld hierboven mag je natuurlijk zowel $4.6 \cdot 10^{-12}$ als 4.6×10^{-12} schrijven. Dat maakt niet uit. Bij grote ronde getallen is het vaak niet duidelijk hoe groot de significantie is. Met de wetenschappelijk notatie kunnen we dit duidelijk maken.

Voorbeeld:

Voorbeeld groot getal Stel dat je het aantal knikkers in een pot hebt geschat op 2500. De onzekerheid is alleen in het laatste getal, maar dat kan je op deze manier niet zien. Je kan dit getal dan beter met de wetenschappelijk notatie schrijven. Bijvoorbeeld: 2.50×10^3 of 25.0×10^2
Op zich mag je ook schrijven 250×10^1 maar in de praktijk doet niemand dit (10^1 gebruiken) en bovendien blijft bij dit voorbeeld dan nog steeds onduidelijk wat de significantie is.

In het algemeen geldt voor de wetenschappelijke notatie het volgende:

- Je schuift de decimale punt op zodat er een getal staat dat in absolute waarde groter is dan 1 en kleiner dan 10. Dit is het getal A.
- Heb je de decimale punt hierbij n plaatsen naar links verschoven dan vermenigvuldig je het getal A met 10^n. Heb je de decimale punt n plaatsen naar rechts verschoven dan vermenigvuldig je A met een factor 10^{-n}.
- Daarna rond je af op het gewenste aantal significante cijfers.

Hieronder een aantal voorbeelden:

Getal	Gewenste significantie	Wetenschappelijke notatie
0.00343	1 cijfer	$3 \cdot 10^{-3}$
0.00343	2 cijfers	$3.4 \cdot 10^{-3}$
0.00343	3 cijfers	$3.43 \cdot 10^{-3}$
10.7	2 cijfers	$1.1 \cdot 10^{1}$
255	2 cijfers	$2.6 \cdot 10^{2}$
34590	2 cijfers	$3.5 \cdot 10^{4}$

Let op! Bij natuurkundige resultaten is het vaak netter om het getal aan te passen aan

een eenheid. Stel dat je een lengte meet, dan kan het netjes zijn om in plaats van 9.2×10^2 meter, 0.92 km te schrijven. De significantie blijft in dit geval hetzelfde. Gebruik de instructies hierboven als richtlijnen en niet als regels. Soms is het beter om ervan af te wijken, maar denk er wel over na!

####Hoeveel significante cijfers noteren?

Het is dus belangrijk om niet te veel en niet te weinig **significante getallen** gebruiken als je een resultaat noteert.

Voor het noteren van een meetresultaat hanteren we de volgende regel:

- Als er **geen** meetonzekerheid op het resultaat bekend is dan noteren we het meetresultaat met 2 significante cijfers.
- Als er **wel** een meetonzekerheid op het resultaat bekend is, dan noteren we de onzekerheid met 2 significante cijfers en noteren we het meetresultaat met dezelfde precisie.

Voorbeeld:

Voorbeeld

Resultaat	Onzekerheid	Notatie
2.515	0.2142	2.52 ± 0.21
2.515	onbekend	2.5
2515	241	$(2.52 \pm 0.24) \cdot 10^3$
2515	onbekend	$2.5 \cdot 10^3$
0.0471	0.12	0.05 ± 0.12
0.00148	10.38	0 ± 10
0.00148	onbekend	0.0015 of $1.5 \cdot 10^{-3}$
24018.2184	1.2125	24018.2 ± 1.2

NB. Als we teruggaan naar het voorbeeld met opmeten van de plank met het meetlint waarbij we hebben gemeten dat de plank 253.3 cm lang is, hebben we 4 significante cijfers genoteerd. Het resultaat is genoteerd zonder meetfout. Toch is dit de juiste notatie geweest. De ingeschatte fout is immers in de orde van een millimeter. In de tabel hierboven wordt steeds aangegeven dat de onzekerheid onbekend is, in zeker zin is die bij de meting van de lengte van de plank *wel* bekend. Meer hierover volgt in de sectie over meetonzekerheid.

####Significantie en berekeningen

Voor het kiezen van het juiste aantal significante cijfers zijn er een aantal regels.

- Bij het vermenigvuldigen of delen van getallen krijgt het resultaat de significantie van het oorspronkelijke getal dat de laagste significantie had.

Voorbeeld:

Vermenigvuldigen we bijvoorbeeld 2.00 (drie significante cijfers) met 3.5 (twee significante cijfers) dan is het resultaat gelijk aan $2.00 \times 3.5 = 7.0$ (twee significante cijfers).

- Bij het optellen of aftrekken van getallen heeft het resultaat niet meer cijfers achter de decimale punt dan het gegeven met het minste aantal cijfers achter de decimale punt.

Voorbeeld:

Tellen we bijvoorbeeld 1.23 op bij 0.1 dan is het resultaat

$$1.23 + 0.1 = 1.3 \tag{12}$$

.

0.4 Datasets visualiseren

1.

In dit deel bekijken we de verschillende manieren om data visueel te presenteren. Aan bod komen grafieken en scatterplots, staafdiagrammen en histogrammen. We laten ook zien hoe je deze met behulp van python kan maken.

Als je data visualiseert dan is het de bedoeling dat iemand anders deze goed kan begrijpen. Er zijn wel een aantal richtlijnen, maar het meest belangrijke is dat de data overzichtelijk is. Dat trends, of juist afwijkingen daarvan, goed zichtbaar worden gemaakt.

De richtlijnen zijn geen regels. Er zijn altijd uitzonderlijke datasets die erom vragen om af te wijken van de richtlijnen. Blijf dus altijd goed nadenken over wat je doet en waarom.

Afhankelijk van wat voor soort metingen je hebt genomen kies je uit een grafiek, een scatterplot, een staafdiagram of een histogram. Elk van deze data visualisatie methodes worden hieronder besproken.

0.4.1 Grafieken & Scatterplots

Grafieken en scatterplots zijn twee vormen van een diagram die veel op elkaar lijken. Ze verschillen wel op een paar punten.

Bij **scattterplots**,

- kunnen voor een ingestelde/gekozen waarde meer dan één gemeten waarden bestaan. Een voorbeeld zou zijn als je een meting doet waarbij je de lengte van mensen opmeet en tegen hun leeftijd uitzet.
- verbind je *nooit* punten met lijnen. Dat zou ook erg verwarrend zijn omdat je de dataset niet altijd logisch kan ordenen. In grafieken is dat vaak trouwens ook onwenselijk.

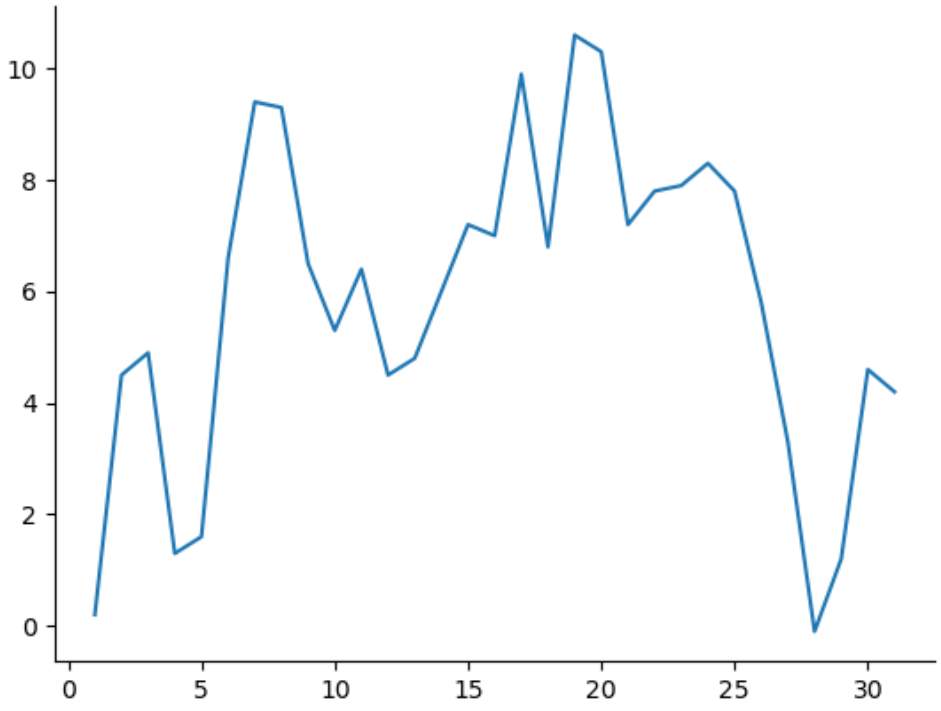
Bij **grafieken**,

- kies je meestal voor een van de twee variabelen een meetpunt of stel je een waarde in. De gemeten waarde laten we zien op de verticale as en de gekozen waarde op de horizontale as.

0.4.1.1 Richtlijnen voor de opmaak van diagrammen

Met behulp van voorbeelden laten we zien wat de richtlijnen zijn en waar je op moet letten.

Stel bijvoorbeeld dat we naar de gemiddelde dagtemperatuur in de maand December 2019 in de Bilt. Hieronder een plot met een lijn tussen elk datapunt (Bron: KNMI, gehomogeniseerde data):

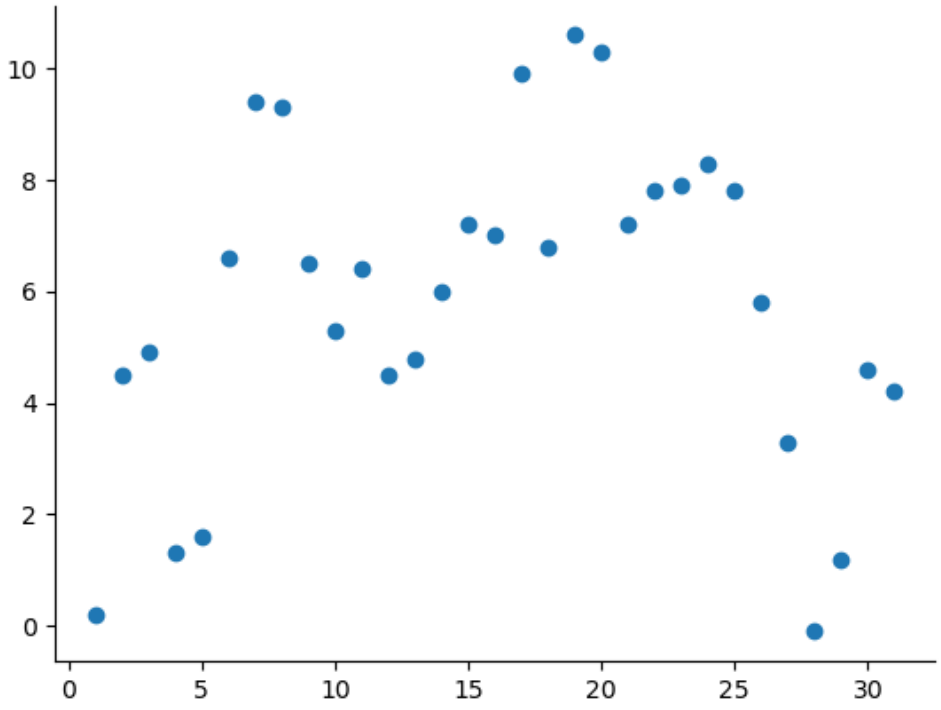


{: width="60%"}

Je ziet dat dit niet erg duidelijk is. Het is bijvoorbeeld niet precies te zien waar de gemeten

punten zitten, we hebben wel een vermoeden voor de plaatsen waarop de lijn abrupt van richting veranderd, maar wie weet zitten er nog wel meer datapunten tussen.

Laten we dezelfde data eens plotten zonder lijnen maar alleen met punten:



{: width="60%"}

Vanuit deze grafiek zien we waar de datapunten zijn. Dat konden we in de lijnplot niet goed zien. We kunnen nu helaas de trend niet meer goed waarnemen. Omdat er op een dag maar één gemiddelde gemeten temperatuur kan bestaan, is het toch beter deze als een grafiek weer te geven. We kiezen ervoor om zowel een lijn als markers te gebruiken.

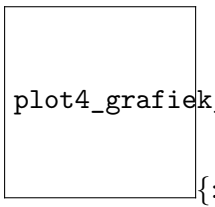
De plot kan echter netter. Zo staan er geen labels op de assen. Nu kunnen we in dit geval wel raden welke as het jaar aangeeft en welke as de temperatuur, maar in veel gevallen is dat niet zo duidelijk. Om die reden moeten er altijd **labels op de assen** staan, zie het figuur hieronder:

`plot3_grafiek_aslabel.png`{: width="100%"}

Zoals je ziet hebben we het formaat van de grafiek ook aangepast zodat de distributie iets natuurlijker overkomt.

Een andere conventie is dat grafieken doorgaans **beginnen bij de oorsprong, tenzij de data dan onvolledig of onleesbaar wordt**. In het geval van het weergeven van de temperaturen wordt de data bijvoorbeeld onvolledig als we de temperatuur bij nul laten beginnen, we hebben immers ook temperaturen onder het vriespunt. In dit geval kunnen we de horizontale as wel bij nul laten beginnen, al is dat voor datums meestal anders.

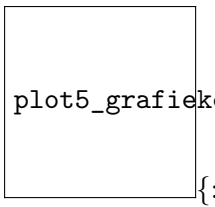
De assen kunnen nog wat netter. Zo eindigt de verticale as net voor de waarde 0, maar het is niet helemaal duidelijk bij welke waarde precies. De horizontale as begint een klein stukje voor 0 en eindigt een klein stukje na 30. Conventie is om assen te laten **beginnen en eindigen op een maatstreepje**. In ons geval laten we het beginnen op de eerste dag van de maand en de laatste dag, daarnaast laten we de temperatuur beginnen op -2 C° en eindigen op 16 C°.



`plot4_grafiek_aslabel_lim.png`

{: width="100%"}

Stel we willen de temperatuur in de Bilt nu weergeven naast de temperaturen gemeten in Vlissingen en Maastricht. De grafiek ziet er dan zo uit:



`plot5_grafieken_aslabel_lim.png`

{: width="100%"}

We hebben ook een legenda toegevoegd zodat duidelijk is welke lijn bij welk weerstation hoort.

Tot nu toe hebben we nog geen titels toegevoegd aan de plots. Dit komt omdat dat voor verslagen en wetenschappelijke artikelen ongebruikelijk is, daar moet het onderschrift namelijk al vertellen wat er te zien is in de grafiek. In webteksten, lesteksten en presentaties kan het echter voorkomen dat een grafiek wel een titel heeft, omdat er in die context vaak geen onderschrift toegevoegd kan worden.

Samengevat:

- Een grafiek van een dataset wordt geplot met punten en eventueel lijnen.
- Het resultaat van een fit of een theoretisch verband wordt met een gladde lijn geplot.

- Bij een enkele dataset wordt geen legenda gebruikt. Als er meerdere datasets in één grafiek worden weergegeven dan is een legenda noodzakelijk.
- Aslabels geven weer wat elke as representeert (inclusief eenheden).
- Assen beginnen in de oorsprong. Een uitzondering kan zijn als de data heel erg ver van de oorsprong af zit.
- Een as begint en eindigt op een groot maatstreepje met een waarde ('major tick') en niet op een klein maatstreepje of een maatstreep zonder getal. Tenzij er een heel goede reden is om hiervan af te wijken. (Zoals in het geval hierboven.)
- Een grafiek voor een wetenschappelijk artikel of een verslag heeft geen titel. Een grafiek voor webteksten of lesmateriaal heeft over het algemeen wel een titel.
- Als je de onzekerheid weet op de variabelen dan is het goed om deze ook weer te geven in je plot. Tenzij deze heel onoverzichtelijk wordt (zoals in een scatterplot met heel veel punten).

Let op! Dit zijn weer richtlijnen en geen regels. Denk altijd goed na over wat je doet en waarom. Het eindresultaat moet goed begrijpbaar zijn en daarvoor is het soms nodig om van de richtlijnen af te wijken.

0.4.2 Staafdiagrammen & Histogrammen

Staafdiagrammen en histogrammen worden allebei typisch gebruikt om frequenties van meetwaardes aan te geven.

Hieronder zie je voorbeelden van een staafdiagram en een histogram.

`autos-in-nederland.png`

{: width="80%"}

Hierboven zie je een **staafdiagram** die de hoeveelheid auto's in Nederland laat zien over drie verschillende jaren opgesplitst naar drie auto categorieën.

`verdeling-van-inkomens-2.png`

{: width="80%"}

Hierboven zie je een **histogram** die de inkomensverdeling in Nederland laat zien.

Er is een belangrijk verschil tussen een staafdiagram en een histogram. Een staafdiagram laat de frequentie zien voor *gecategoriseerde* verdelingen. Een histogram wordt gebruikt om het resultaat van een *numeriek sorteerbare* verdeling mee weer te geven. In het geval

van een histogram gaat het vaak om data met een continue variabele, zoals bijvoorbeeld bij het opmeten van lengte of gewicht. In dat geval sorteer je de data per interval.

Bij het weergeven van data in een histogram wordt de data gegroepeerd in intervallen. De breedte van de staven (in het vervolg 'bins' genoemd) geeft de breedte van de intervallen.

Bij een staafdiagram kun je de frequentie direct aflezen; voor één categorie lees je op de as af hoe vaak deze voorkomt. Voor een histogram is de frequentie gelijk aan de oppervlakte van de balken, en dus afhankelijk van de bin breedte.

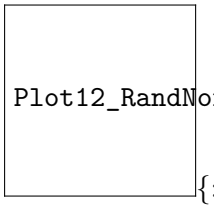
0.4.2.1 Breedte van de bins bij een histogram

Voor een histogram is de breedte van de intervallen van belang. Als we te weinig bins kiezen dan worden de intervallen erg groot (/breed) en is er minder te zeggen over het gedrag van de data. Als we te veel bins kiezen dan fluctueert de hoogte van de (smalle) bins onderling erg en is het ook lastiger om de trend in de data goed in te schatten.

Dit bekijken we aan de hand van een voorbeeld. Zo zou het kunnen zijn dat het ideale plaatje bij een gegeven dataset het volgende is:

{: width="400px"}

Als we te brede bins kiezen dan wordt de data afgevlakt en kunnen we het bovenstaande gedrag niet meer herkennen:



{: width="400px"}

Kiezen we juist te smalle bins, dan kunnen we het gedrag van de data nog wel herkennen (in dit geval) maar er is veel fluctuatie in de hoogte van de bins:

{: width="400px"}

Met het kiezen van te veel bins hebben we dus visuele ruis geïntroduceerd, dit maakt het moeilijker om het gedrag op het oog te herkennen.

Bij het bepalen van het optimale aantal bins en de optimale bin breedte is het belangrijkste dat het gedrag van de data goed zichtbaar is. Er zijn verschillende formules (bijvoorbeeld de square of de Sturges formule) ontwikkeld waarmee je het aantal bins dat je nodig hebt kunt berekenen. Echter, geen van die formules kun je blind toepassen. Het is veel beter om gewoon goed naar je dataset te kijken en een inschatting te maken van de bin breedte.

Bij het maken van een histogram moet je goed letten op het volgende:

- Het bereik (de range) die je kiest op de horizontale as. Van waar tot waar plot je de data? Meestal wil je de gehele dataset laten zien, maar soms wil je juist inzoomen op een kleiner stukje.
- De bin breedte. Meestal kies je voor het hele histogram dezelfde bin breedte, in sommige gevallen kun je verschillende bin breedtes kiezen. In elk geval geldt dat het histogram goed 'leesbaar' moet zijn. Het moet duidelijk blijven hoe de data gedistribueerd is. Wat is de trend? Zijn er afwijkingen van die trend.
- Let bij histogrammen erg goed op waar de grens van een bin ligt. Vooral als je een dataset met natuurlijke getallen weergeeft is het belangrijk dat de bin grenzen netjes *tussen* de natuurlijke getallen ligt. Anders kan de distributie van de data verkeerd gerepresenteerd worden.
- Het histogram is makkelijker leesbaar als de bins een natuurlijk interval hebben. Als je range van 0 tot 10 loopt is het heel gek om deze te verdelen in 7 bins.

Voor zowel histogrammen als staafdiagrammen geldt:

- Natuurlijk moeten bij histogrammen en staafdiagrammen ook netjes aslabels worden gebruikt.
- Als je meer dan één dataset laat zien maak dan gebruik van een legenda.

0.4.3 Wanneer gebruik je wat?

1. Als je de incidentie (of frequentie) van meetwaarden wil laten zien dan gebruik je een histogram of staafdiagram.
   - Een staafdiagram gebruik je als de meetwaarden discreet zijn gecategoriseerd, bijvoorbeeld in het soort auto of per kleur.
   - Een histogram gebruik je voor variabelen die numeriek geordend kunnen worden, zoals bijvoorbeeld variabelen met integer of continue waardes.
2. Als je de relatie tussen twee variabelen wilt tonen kies je voor een grafiek of een scatterplot.
   - Je gebruikt een grafiek als de afhankelijke variabele (meestal de langs de x-as) unieke waardes kent. Dus voor een bepaalde waarde van x is maar één uitkomst van y mogelijk. Andersom zijn er wellicht meerdere waardes voor x voor een bepaalde gemeten grootheid y. Bijvoorbeeld de gemeten temperatuur om 12 uur 's middags op een bepaalde locatie. Er kan maar 1 gemeten temperatuur bestaan.

- Je gebruikt een scatterplot als er geen unieke waarde is per afhankelijke variabele. Bijvoorbeeld als je de lengte van een student meet in relatie met de leeftijd. Er zijn waarschijnlijk meerdere studenten met dezelfde leeftijd in de groep die hoogstwaarschijnlijk in lengte van elkaar verschillen.

0.4.4 Data plotten met Python

Om in Python te kunnen plotten moeten we als eerste een library importeren die ingebouwde functies heeft voor het visueel weergeven van data. Een populair pakket is Matplotlib, deze zullen we in dit vak dan ook gebruiken (er zijn ook andere geschikte pakketten zoals Seaborn, geplot en Plotly). We importeren de `pyplot` functie vanuit Matplotlib en geven deze de naam 'plt' met het volgende commando:

```
import matplotlib.pyplot as plt
```

De naamgeving `plt` met het commando `as plt` is optioneel, maar wel handig omdat we deze functie over het algemeen vaak zullen gebruiken (dat scheelt typen).

####Voorbeeld: een grafiek plotten Stel we hebben de hoogte van een vallende bal gemeten als functie van de tijd. In de tabel hieronder is de gemeten data weergegeven:

t (s)	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
h(cm)	180.0	178.8	175.1	169.0	160.4	149.3	135.9	120.0	102.0	80.7	57.4	31.6	3.4

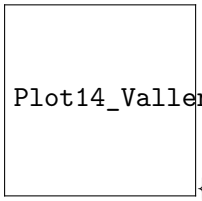
Nu maken we een lijst `t_data` aan voor de tijd en een lijst `h_data` voor de hoogte van de bal:

```
t_data = [0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0]
h_data = [180.0, 178.8, 175.1, 169.0, 160.4, 149.3, 135.9, 120.0, 102.0, 80.7, 57.4, 31.6, 3.
```

Daarna roepen we het `plot` commando uit matplotlib.pyplot aan:

```
plt.plot(t_data, h_data, 'ro')
```

Met `'ro'` geven we aan dat we rode gevulde punten in de plot willen. De plot ziet er nu als volgt uit:



{:width="400px"}

Je ziet dat de assen automatisch vanaf de laagste waarde tot aan de hoogste waarden gaan, en hierbij niet eindigen op een maatstreepje. Daarnaast willen we graag labels op de assen.

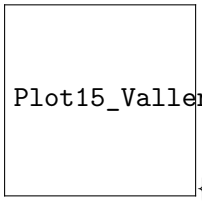
De limiet van de assen kunnen we aangeven met de commando's `plt.xlim` en `plt.ylim`:

```
plt.xlim(0,7)
plt.ylim(0,200)
```

Labels voor de assen kunnen we als volgt specificeren:

```
plt.xlabel('t (s)')
plt.ylabel('h (cm)')
```

Het resultaat is:



Plot15_Vallenbal_PlotOpgemaakt.png

{: width="400px"}

De volledige code tot nu toe is:

```
## dataset in lijsten zetten
t_data = [0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6]
h_data = [180,178.8,175.1,169.0,160.4,149.3,135.9,120,102,80.7,57.4,31.6,3.4]

## data plotten, as-limieten instellen, as-labels instellen
plt.plot(t_data, h_data, 'ro')
plt.xlim(0,7)
plt.ylim(0,200)
plt.xlabel('t (s)')
plt.ylabel('h (cm)')
```

Als we nu nog een dataset hebben, bijvoorbeeld van dezelfde bal die vanaf een hoogte van 160 cm valt in plaats van een hoogte van 180 cm:

```
## tweede dataset in lijsten zetten
t_data2 = [0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5]
h_data2 = [160,158.8,155.1,149.0,140.4,129.3,115.9,100,82,60.7,37.4,11.6]
```

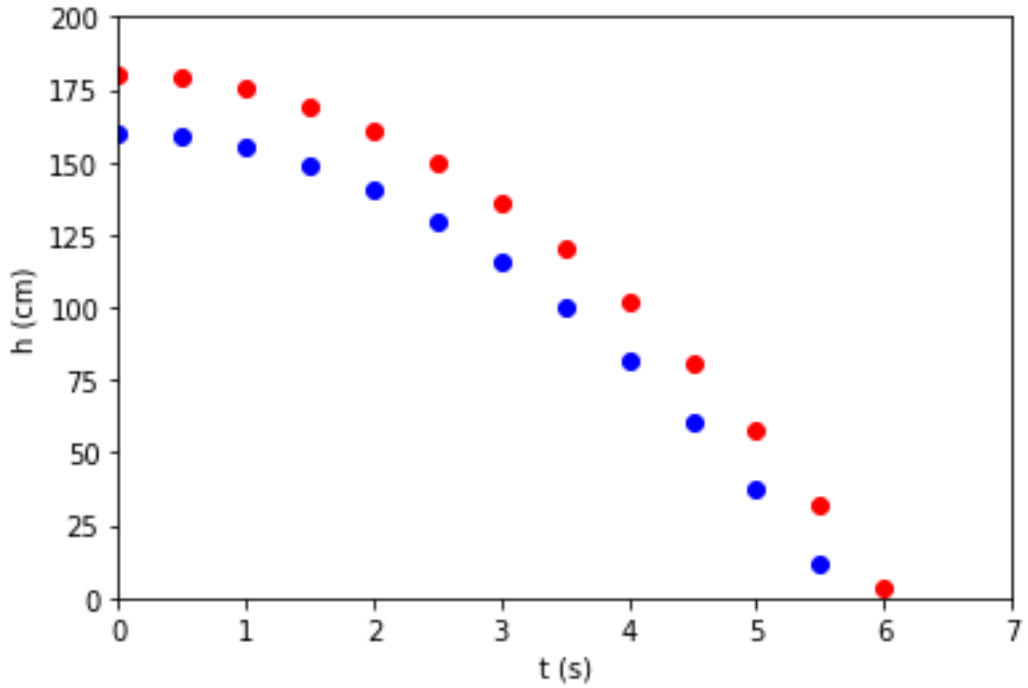
Deze dataset kunnen we in de grafiek van de eerste plotten door twee keer het plot commando achter elkaar te gebruiken: daarna gebruiken we weer dezelfde eigenschappen voor de limieten en de aslabels:

```
plt.plot(t_data, h_data, 'ro')
plt.plot(t_data2, h_data2, 'bo')
```

Daarna gebruiken we weer dezelfde eigenschappen voor de as-limieten en de as-labels":

```
plt.xlim(0,7)
plt.ylim(0,200)
plt.xlabel('t (s)')
plt.ylabel('h (cm)')
```

De plot ziet er dan als volgt uit:



{:
width =“400px”}

Omdat er meerdere datasets in één grafiek zijn weergegeven is het noodzakelijk om hier een legenda bij te plaatsen. Een legenda kan op meerdere plaatsen in de figuur neergezet worden. Voordat we de legenda kunnen toevoegen moeten we de plots eerst labelen dit doen we door `label = "naam"` achteraan in de `plot` commando's toe te voegen:

```
plt.plot(t_data, h_data, 'ro' , label='h(0) = 180 cm')
plt.plot(t_data2, h_data2, 'bo', label='h(0) = 160 cm')
```

Nu kunnen we de legenda als volgt toevoegen (hier kiezen we ervoor om de legenda in de rechterbovenhoek neer te zetten zodat er geen overlap is met de grafieken zelf):

```
plt.legend(loc='upper right', shadow=True, ncol=1)
```

De grafiek is nu als volgt:

PLOT18_MeerdereScatter_legenda.png

{: width =“400px”}

De volledige code tot nu toe is:

```
## dataset in lijsten zetten
t_data = [0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6]
```

```
h_data = [180,178.8,175.1,169.0,160.4,149.3,135.9,120,102,80.7,57.4,31.6,3.4]

## tweede dataset in lijsten zetten
t_data2 = [0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5]
h_data2 = [160,158.8,155.1,149.0,140.4,129.3,115.9,100,82,60.7,37.4,11.6]

## data plotten, as-limieten instellen, as-labels instellen
plt.plot(t_data, h_data, 'ro' , label='h(0) = 180 cm')
plt.plot(t_data2, h_data2, 'bo', label='h(0) = 160 cm')
plt.xlim(0,7)
plt.ylim(0,200)
plt.xlabel('t (s)')
plt.ylabel('h (cm)')

## legenda toevoegen
plt.legend(loc='upper right', shadow=True, ncol=1)
```

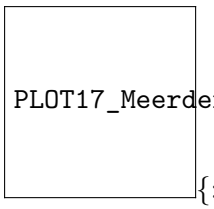
Een ander voorbeeld is dat we lijnen willen plotten, bijvoorbeeld van een theoretisch verband. (De conventie is dat data altijd met punten wordt uitgebeeld.) Dit kan je als volgt doen:

```
## datasets in lijsten
x = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y1 = [1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21] ##2x+1
y2 = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] ##x+2

## datasets plotten als 'solid' lijnen
plt.plot(x, y1, 'r-' , label='dataset 1')
plt.plot(x, y2, 'b-', label='dataset 2')
plt.xlim(0,7)
plt.ylim(0,20)
plt.xlabel('x')
plt.ylabel('y')

## legenda toevoegen
plt.legend(loc='upper right', shadow=True, ncol=1)
```

De bijbehorende plot:



{: width="400px"}

####Voorbeeld: een histogram plotten In de allereerste opgave M1.1 ga je een histogram plotten. In die opgave staat stap voor stap uitgelegd hoe je dat moet doen.

Maar kijk vooral ook in de online manual van matplotlib.

Succes!