

Onderwijs & Communicatie

Ian de Ronde, Nishant Mangre, Emma Holopainen

1 juli 2023

1 Statistiek met de computer

Wiskunde is een heel handig gereedschap om conclusies te trekken uit een onderzoek. Zo worden veel belangrijke dingen berekend, bijvoorbeeld klimaatverandering. Dit wordt berekend door over een lange periode te meten hoe het weer is, bijvoorbeeld elke dag voor vijftig jaar lang. Het probleem is dat je hierna duizenden metingen hebt voor het weer. Dit is een enorme data set en daaruit kan je moeilijk de belangrijke patronen zien.

Het doel van statistiek is om zulke grote data sets samen te vatten tot de belangrijkste conclusies. Bij onze weermetingen zou een mogelijke conclusie zijn dat de temperatuur over de jaren is gestegen. Er zijn ook andere berekeningen die mogelijk interessant kunnen zijn, bijvoorbeeld het gemiddelde, de meest voorkomende temperatuur (de mediaan) en voorspelbaarheid van het weer.

Dit is moeilijk met de hand te berekenen, dus gebruiken we de computer. Er zijn verschillende programma's om statistische berekeningen te maken. Wij gebruiken nu Google Spreadsheets. Dit is een programma van Google en wordt ook gebruikt door onderzoekers en bedrijven.

Om te beginnen open Google spreadsheets:

Hiervoor moet je eerst Google-drive openen.

Druk vervolgens op Nieuw, links bovenin.

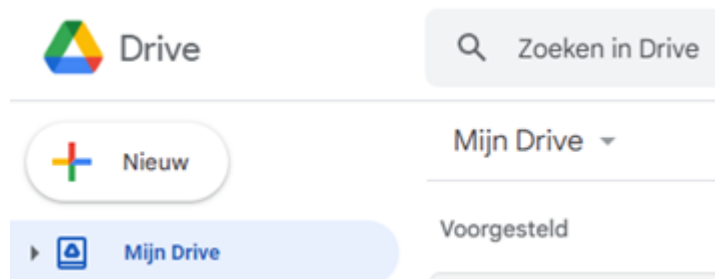
Druk daarna op Google spreadsheets om een leeg spreadsheets bestand te maken.

Nu zit je in een spreadsheets bestand. In de vakjes kan je getallen zetten en vervolgens berekeningen en tabellen maken.

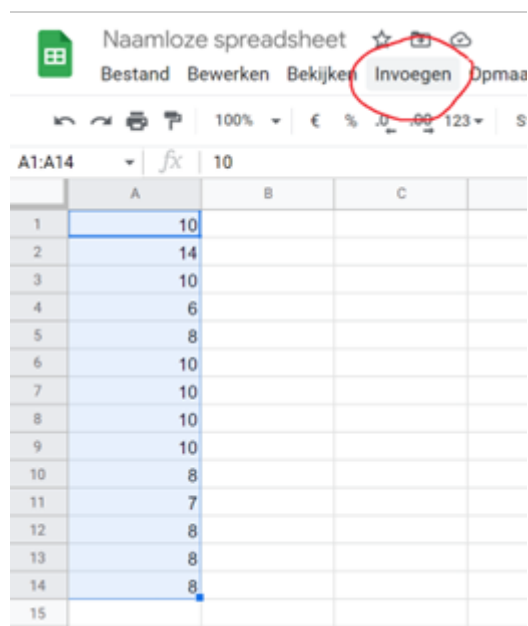
Opgave 1.1 (Data invoeren). *Zoek het weer op voor de komende twee weken en type de verwachte temperatuur per dag onder kolom A.*

Opgave 1.2 (Diagram maken). *Maak vervolgens een grafiek van de temperatuur voor de komende twee weken. Dit doe je als volgt:*

- *Selecteer alle getallen onder kolom A.*
- *Druk op de knop invoegen, deze staat bovenaan.*
- *Druk daarna op diagram*



Figuur 1: Beginscherf van Google Drive



Figuur 2: Google Spreadsheets

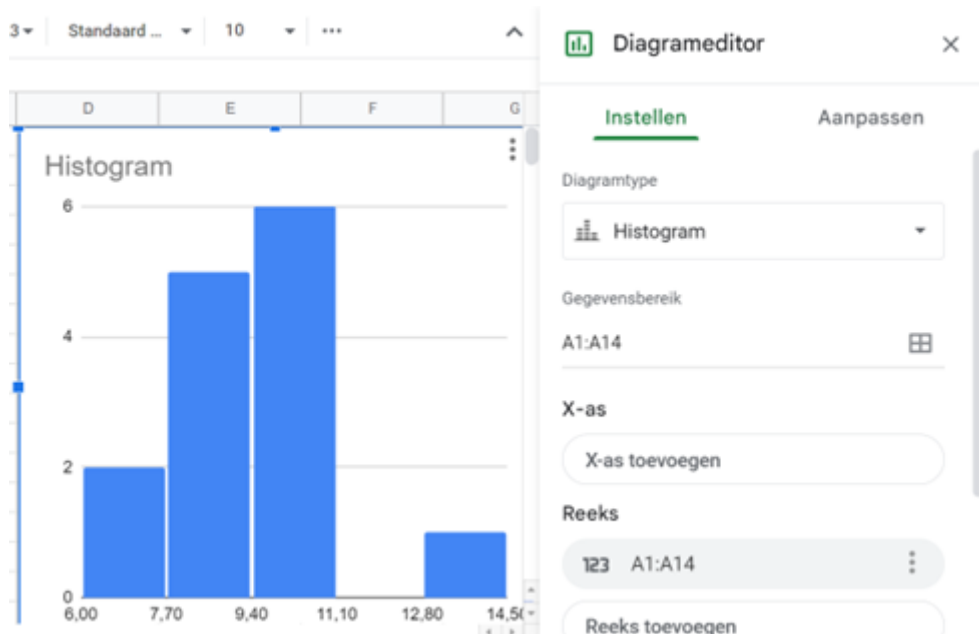
Nu opent er een histogram en de diagrameditor. Verander in de diagrameditor het type diagram naar een lijn.

Opgave 1.3 (Mooie diagrammen maken). *Om je resultaten duidelijk te maken, wil je dat je grafiek er mooi uitziet.*

Open weer de diagrameditor en klik op aanpassen. Verander een paar instellingen zodat je grafiek er anders uitziet. Geef je grafiek sowieso een titel.

Om onze grafiek nog duidelijker te maken kunnen we ook de datums van de temperaturen toevoegen op de x-as. Dit kunnen handmatig doen, maar er is ook een handige truc waardoor dit veel sneller kan. Dit laten we zien in de volgende opdracht.

Opgave 1.4 (Vulgreep). *Nu is het de opdracht om de datums toe te voegen van de temperaturen. Maar dit gaan we niet handmatig doen. Met behulp van de vulgreep kan de computer de datum voor een gehele kolom invullen.*



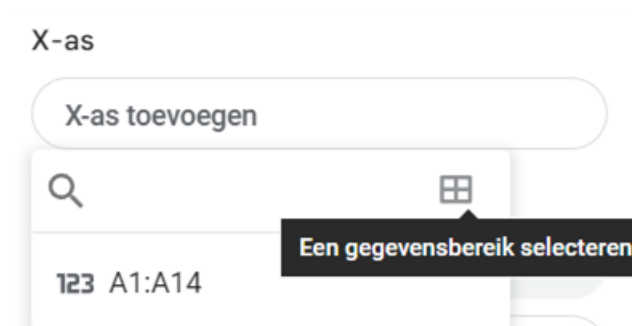
Figuur 3: De diagrammeditor

- Vul ten eerste zelf te datum in van de eerste dag. Zet dit in kolom B, naast de eerste temperatuur.
- Selecteer vervolgens het vakje van de datum. De vulgreep is het blauwe vakje recht onderin.
- Sleep de vulgreep naar onderen, zodat elke temperatuur een datum ernaast heeft.

Met de vulgreep kan je een heleboel tijd besparen. Maar soms vult het niet precies in wat je zou willen hebben.

Vul zelf een paar getallen in en kijk wat de vulgreep daarna invult. Voeg nu de datum toe aan de grafiek.

- Open de diagrammeditor en druk op X-as toevoegen.
- Druk op 'een gegevensbereik selecteren'. Dan opent het menu waarmee we aangeven waar de datums staan.
- In het menu kunnen we de vakjes invullen of we selecteren de vakjes met de muis.
- Probeer beide om de datums op de x-as te krijgen.



Figuur 4: Vernader het gegevensbereik

Probeer nu de overige instellingen in de diagrameditor. Houdt de instellingen als de grafiek hier mooier van wordt.

Google Spreadsheets slaat automatisch je voortgang op, dus dat hoeft je niet zelf te doen. Het enige probleem is dat je bestand wordt opgeslagen onder de naam 'Naamloze spreadsheet'. Wanneer je tevreden bent met je grafiek, verander je de naam van je bestand zodat je het makkelijk kunt terugvinden.

2 Berekeningen in Spreadsheets

Google spreadsheets heeft ook een ingebouwde rekenmachine. Deze is heel handig als je berekeningen wilt doen met data. Een berekening kan in een vakje. Selecteer een vakje en type vervolgens: $= 5 + 2$
Daarna komt in het vakje het antwoord van de berekening te staan.

Voor een berekening moet altijd een $=$ staan.

We kunnen ook berekeningen doen met de data uit ons spreadsheets bestand. Hiervoor hebben we eerst wat data nodig.

Open de volgende link:

<https://opendata.cbs.nl/statline//CBS/nl/dataset/85184NED/table?dl=7A888>

Hierin staat het aantal jongeren dat afgestudeerd is in een bepaalde richting. Er staan negen verschillende richtingen in de tabel tussen de jaren 2003 en 2021.

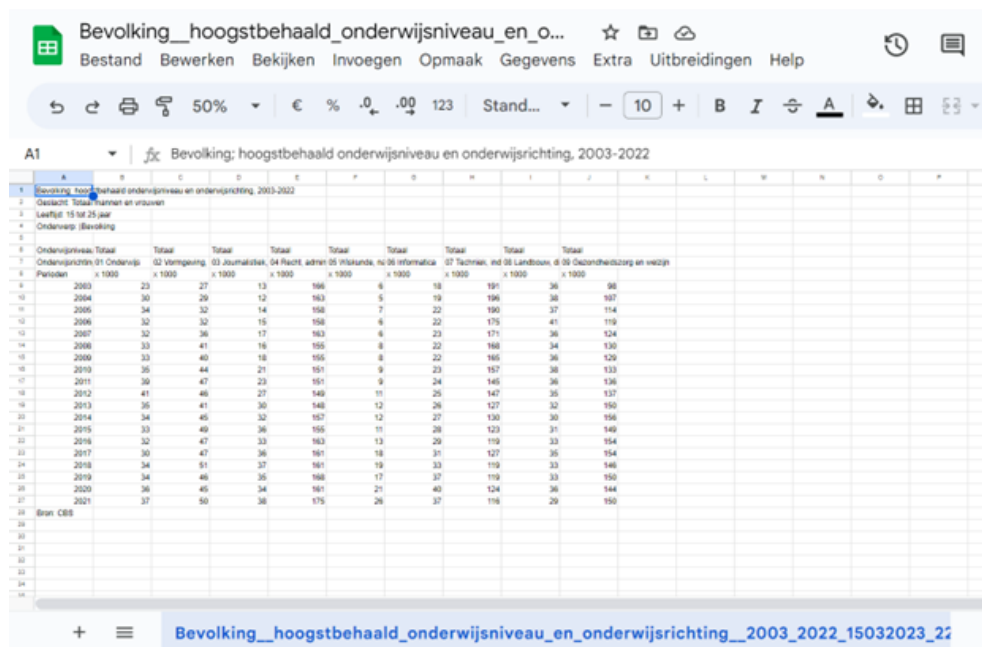
- Download vervolgens het bestand als 'CSV volgens tabelindeling'.



Figuur 5: Een tabel van het CBS

- Nu moeten we het bestand importeren in Google Spreadsheets. Open hiervoor eerst Google Drive.
- Klik vervolgens op 'Nieuw' en upload de data set.
- Nadat je dit hebt gedaan kan je de tabel ook in Google Spreadsheets openen.

Zo ziet het bestand eruit in Spreadsheets:



Figuur 6: Een tabel van CBS in Google Spreadsheets

Opgave 2.1 (Berekeningen in Spreadsheets). *De tabel is bevat veel informatie over studierichtingen. Maar wat wel mist zijn het totale aantal mensen dat überhaupt studeert.*

Zoek in de tabel voor elke studierichting het aantal afgestudeerden in het jaar 2021. Kies vervolgens een leeg vakje. Bereken hierin het aantal afgestudeerden in het jaar 2021 voor alle richtingen bij elkaar opgeteld.

Het handige van rekenen in spreadsheets is dat we ook meteen de data kunnen selecteren voor een berekening.

Voorbeeld 2.1 (Berekeningen met vakjes). *Stel je voor dat we willen berekenen hoeveel mensen zijn afgestudeerd in 2003 in een richting waarvoor biologie nodig is. Dit zijn richtingen 8, Landbouw, diergeneeskunde en -verzorging, en richting 9, Gezondheidszorg en welzijn.*

Zoek in welke vakjes het aantal afgestudeerde voor beide richtingen staan. Zoek ook in welke rij het jaar 2003 staat.

We zien dat richting 8 onder kolom I staat en dat richting 9 onder kolom J staat. Het jaar 2003 staat in rij 9.

Het vak waarin het aantal afgestudeerde in richting 8 in het jaar 2003 staat noemen we dan I9. Het vak waarin het aantal afgestudeerde in richting 9 in het jaar 2003 staat noemen we dan J9.

Om de som van deze twee getallen te krijgen kunnen we simpelweg: =I9 + J9, invoeren in een leeg vakje.

14 fx =I9+J9

	F	G	H	I	J	K
2						
3						
4				134		
5						
6	Totaal	Totaal	Totaal	Totaal	Totaal	
7	05 Wiskunde, nat	06 Informatica	07 Techniek, ind	08 Landbouw, ind	09 Gezondheidszorg en welzijn	
8	x 1000	x 1000	x 1000	x 1000	x 1000	
9	6	18	191	36	98	
10	5	19	196	38	107	

Figuur 7: In I4 is de berekening gedaan

Opgave 2.2 (Berekeningen met vakjes). *Voer nu dezelfde berekening uit als in opdracht 5, maar dan door de vakjes te gebruiken.*

Bereken ook het totaal aantal afgestudeerde in het jaar 2003. Doe dit in het vak L9.

Om het totaal aantal afgestudeerde voor elk jaar te bepalen kunnen we de vulgreep gebruiken. Selecteer het vak L9 en sleep deze door om het totaal aantal afgestudeerde voor elk jaar te berekenen.

Onze nieuwe kolom bevat nu gegevens die we in de vorige tabel niet konden vinden. Kies een passende naam voor deze kolom en type dit in het vak L7, zodat het mooi past bij de rest van de tabel.

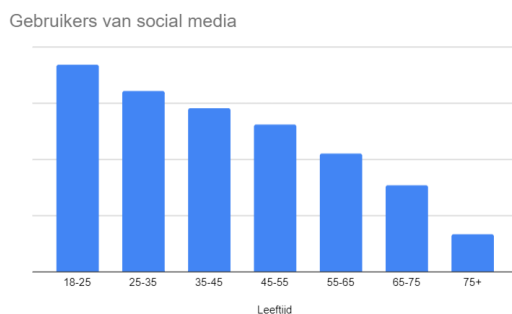
Als je in het vak L7 een lange titel hebt getypt kan het zijn dat die over een aantal ander vakjes gaat. Maar let op dat de titel alsnog alleen in het vak L7 zit. De andere vakjes zijn nog leeg. Als je een waarde geeft aan deze lege vakjes, dan is een deel van de titel van L7 verborgen. Dit verandert echter niks aan wat erin L7 staat.

L7 ▼ fx Totaal aantal afgestudeerde				
	K	L	M	N
4				
5				
6				
7	zorg en welzijn	Totaal aantal a	23	

Figuur 8: De titel van L7 ging eerst ook over M7 heen. Maar door M7 een waarde te geven is dit verdwenen. We zien bovenin nog steeds dat L7 de volledige titel als waarde heeft.

3 Verdelingen maken in Spreadsheets

We kunnen met Google Spreadsheets ook een verdeling maken uit een tabel. Dit doen we door een histogram te maken van gemeten waarden. Een histogram is namelijk veel overzichtelijker dan een tabel (zie figuur 9).



Figuur 9: Een histogram van de leeftijden van de gebruikers van social media (bron: cbs). Aan de hand het histogram zien we meteen dat jongeren mensen meer social media gebruiken dan oudere mensen. Hiervoor hoeven dus niet alle metingen te kennen.

Opgave 3.1 (a). (Een verdeling maken uit een frequentie tabel.)

Bekijk de volgende tabel:

Leeftijdsklasse	Frequentie klinische opname
0	9909,7
1-5	506,9
5-10	220,9
10-15	230,0
15-20	409,5
20-25	586,6
25-30	940,7
30-35	1071,7
35-40	809,0
40-45	675,3
45-50	776,0
50-55	964,5
55-60	1197,8
60-65	1508,8
65-70	1922,8
70-75	2454,3
75-80	3077,8
80-85	3431,2
85-90	3593,2
90-95	3475,9
95+	2827,5

Dit zijn het aantal klinische ziekenhuisopnames uit 2012 (bron: cbs). De aantallen staan per 10 000 mensen in de bevolking.

Dus in 2012 zijn per 10 000 mensen die 35-40 jaar oud is, 809 klinisch opgenomen. Aan de hand van deze data kunnen we bepalen welke leeftijdsklassen een groter risico heeft op ziekenhuisopname. Dit gaan we doen door te kijken naar een grafiek van de verdeling.

- Kopier eerst de data in een Spreadsheet bestand.
- Selecteer de data en voeg een diagram in.
- Geef het diagram een titel

Een klinische opname is een ziekenhuisopname die meerdere dagen duurt. Voorbeelden van klinische opnames zijn: Uitgebreid onderzoek, langdurige behandeling en bevalling van een moeder.

(b) Wat valt je op aan de grafiek? Welke leeftijdsklasse worden het vaakst opgenomen, en om wat voor reden zijn deze mensen opgenomen?

(c) Bereken de gemiddelde leeftijd van mensen die opgenomen worden.

Herriner dat de formule voor het gemiddelde gelijk is aan

$$\text{Gemiddelde} = \frac{\text{Som van de waarnemingen}}{\text{Som van het aantal waarnemingen}}$$

De waarnemingen zijn de leeftijden van opgenomen patiënten.

We kunnen dit berekenen in Google Spreadsheets. Eerst berekenen we de som van de leeftijden.

- Schrijf eerst naast elke leeftijdsklasse de gemiddelde leeftijd.
- Dan vermenigvuldigen we de leeftijd met het aantal opnamen. Doe de berekening in het vakje.
- Doe dit voor elke leeftijdsklasse. Dit kan met de vulgreep.

	A	B	C	D
1	Leeftijdsklasse	Aantal opnames	Gemiddelde leeftijd	Aantal opnames * gemiddelde leeftijd
2	0	9909,7	0	=B2*C2
3	1-5	506,9	2,5	=B3*C3
4	5-10	220,9	7,5	=B4*C4
5	10-15	230		
6	15-20	409,5		
7	20-25	586,6		
8	25-30	940,7		
9	30-35	1071,7		
10	35-40	809		
11	40-45	675,3		
12	45-50	776		
13	50-55	964,5		
14	55-60	1197,8		
15	60-65	1508,8		
16	65-70	1922,8		
17	70-75	2454,3		
18	75-80	3077,8		
19	80-85	3431,2		
20	85-90	3593,2		
21	90-95	3475,9		
22	95+	2827,5		

Hierboven is een voorbeeld voor de eerste drie leeftijdsklassen.

De kolom 'Gemiddelde leeftijd' is zelf berekend, terwijl de kolom 'Aantal opnames * gemiddelde leeftijd' door Spreadsheets is berekend. Wanneer je op enter drukt dan komt het antwoord van de berekening te voor schijn.

Deze tabel kan je aanvullen voor de rest van de leeftijdsklassen.

- Neem de som van alle waarde uit de kolom 'Aantal opnames * gemiddelde leeftijd'.
- Bereken het totaal aantal opnames.
- Bereken nu het gemiddelde met behulp van de formule.

(d) Je hebt nu het gemiddelde berekend. Hebben mensen van deze leeftijd een grote kans om opgenomen te worden in het ziekenhuis?

Cummulatieve verdelingen

Hieronder zie je een gewone verdeling voor het aantal mensen dat op een dag naar de Efteling gaat. Dit is een schatting voor de maand Juli.



Aan de hand van deze gegevens wilt de Efteling bepalen hoeveel personeel ze nodig zullen hebben in de maand Juni.

Opgave 3.2. (Een cummulatieve verdeling maken.)

(a) Hoeveel bezoekers zal de Efteling het vaaksts krijgen per dag? En hoeveel bezoekers het meest op een dag?

De Efteling wilt zoveel personeel zodat er precies genoeg is voor het aantal bezoekers aanwezig op die dag. Maar van te voren weten ze niet exact hoeveel mensen er aanwezig zullen zijn.

(b) Waarom is het niet verstandig om personeel in te huren aan de hand van de aantallen bepaald bij (a).

De Efteling wilt dat er minstens 90% van de dagen er genoeg personeel aanwezig is.

(c) Hoeveel dagen moet er dan genoeg personeel aanwezig zijn.

Om te bepalen hoeveel personeel genoeg is voor minstens zoveel dagen hebben we een cummulatieve verdeling nodig.

Aantal bezoekers (in duizendtallen)	Frequentie
20	3
21	3
22	6
23	3
24	4
25	2
26	1
27	1
28	2
29	1
30	2
31	1
32	0
33	0
34	0
35	1

Dit is een tabel van de verdeling.

(d) Neem deze over in Google Spreadsheets.

Hiervan gaan we een cummulatieve verdeling maken.

- Maak een nieuwe kolom voor de cummulatieve waarden.
- Voor elk aantal bezoekers is dit gelijk aan het aantal dagen dat er zoveel bezoekers of minder zijn gekomen.
Dus de cummulatieve waarde voor 23 duizend bezoekers is gelijk aan: De som van de frequenties van het aantal bezoekers voor 20, 21, 22 en 23 duizend.

	A	B	C
1	Aantal bezoekers	Frequentie	Cummulatieve waarde
2	20	3	=SOM(B2:B2)
3	21	3	=SOM(B2:B3)
4	22	6	=SOM(B2:B4)
5	23	3	
6	24	4	
7	25	2	

Hier is een voorbeeld voor de eerste drie waarden.

- Wanneer je dit voor alle waardes hebt gedaan, maak een diagram van de cummulatieve waarden tegenover het aantal bezoekers.

De helling van een cummulatieve verdeling zegt hoeveel dagen er zijn waar het nodig is om meer personeel aan te nemen. Dus is het verstandiger om personeel aan te nemen wanneer de helling steil is.

(e) Wanneer neemt de helling af van je cummulatieve verdeling? Voor hoeveel bezoekers zou je personeel aan nemen?

Steekproef

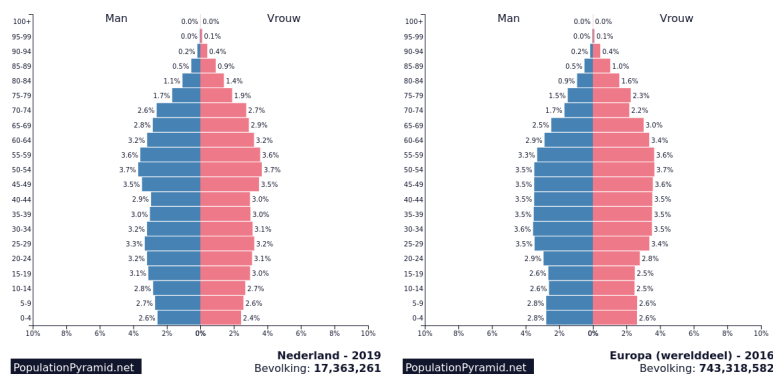
Als je een verdeling wilt maken van een groep dan moet je eerst data verzamelen. Dit kan door bijvoorbeeld metingen en enquêtes. Maar dit proces kan lang duren, vooral als je het voor een hele grote groep mensen moet doen.

Bij een **steekproef** dan meet je alleen een klein deel van de groep. Dan is het makkelijker om te meten en een verdeling te maken. Het gebruik van een steekproef moet wel veroorloofd zijn, omdat je niet iedereen meeneemt in de verdeling.

Om dit te voorkomen is het belangrijk dat de steekproef **representatief** is. Dit betekent mensen uit de steekproef dezelfde kenmerken hebben als mensen van de gehele groep. Zo komt elke bevolkingsgroep voor in de verdeling.

Voorbeelden:

Als je een steekproef wilt nemen van de nederlandse bevolking, dan kan je steekproef niet alleen uit Amsterdam komen. In dit geval hou je geen rekening met mensen uit dorpen. Dus is deze steekproef niet representatief voor heel Nederland.



Figuur 10: Verdelingen van de bevolkingsleeftijd voor Nederland en Europa.

In het bovenstaande figuur staan twee verschillende verdelingen. Voor de verdeling van Europa is een ontzettend grote meting gedaan. Maar we kunnen aannemen dat de verdeling van leeftijd in Nederland ongeveer hetzelfde is als in Europa. Dus zouden we ook Nederland kunnen gebruiken als steekproef. Dan hoeven we niet mensen uit heel Europa te meten.

Opgave 3.3. (Verdeling maken uit een steekproef.)

Jullie klas is een goeie steekproef voor middelbarenscholieren uit Hoorn.

(a) Kies een onderwerp dat je wilt meten en meet dit in je klas. Reistijd naar school, lengte en telefoongebruik zijn enkele voorbeelden.

(b) Maak eerst een frequentie tabel van de metingen.

- Maak eerst een tabel van je metingen in Google Spreadsheets.

- Bepaal ongeveer acht waardes die al je waardes opsplitsen. Maak hiermee een nieuwe kolom.

Metingen lengte	Tussenwaardes
168	150
175	160
162	170
177	180
184	190
192	200
188	
165	
172	
178	

Hierboven staat een voorbeeld voor de lengte. Links staan de metingen en rechts de tussenwaarde. De tussenwaarde splitsen de mogelijke lengtes op in intervallen van 10cm.

- De functie '=INTERVAL' kan uit deze gegevens een frequentie tabel maken. Geef deze functie de metingen en de intervallen.

	B	C	D	E
38		Metingen lengte	Tussenwaardes	Frequentie tabel
39		168	150	=INTERVAL (C39:C48;D39:D44)
40		175	160	0
41		162	170	3
42		177	180	4
43		184	190	2
44		192	200	1
45		188		0
46		165		
47		172		
48		178		
49				

Hierboven staat een voorbeeld. Uit de tabel kunnen we aflezen dat er 4 metingen zijn van lengtes tussen de 170-180cm en 0 metingen van lengtes van 200cm of groter.

- (c) Maak een diagram van de frequenties tegenover de tussenwaardes.
- (d) Denk je dat deze verdeling representatief is voor middelbarenscholieren uit heel Nederland?
Denk je dat deze verdeling representatief is voor de gehele bevolking uit Hoorn?
Geef een uitleg voor beide vragen.