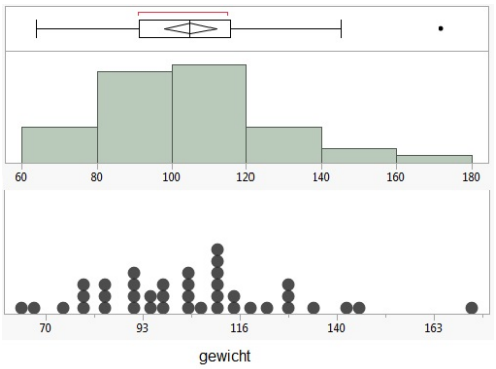
2.1 Data presenteren

Inleiding

Een van de belangrijkste zaken in de statistiek is het goed ordenen van je gegevens, je data. Dat komt omdat je om goede uitspraken te kunnen doen, meestal met grote hoeveelheden gegevens moet werken. En over zo'n grote brij aan gegevens verlies je snel het overzicht. Het werken met overzichtelijke figuren als dotplots en staaf- of lijndiagrammen en het indelen in klassen is dan nuttig.



Figuur 1

Je leert in dit onderwerp

- de begrippen discrete en continue variabelen en deze te onderscheiden;
- de data in dotplots en staafdiagrammen te interpreteren;
- de begrippen klassenbreedte en klassengrens.

Voorkennis

- de begrippen data, populatie, steekproef, aselect en representatief, kwantitatief en kwalitatief, absolute en relatieve frequentie;
- statistische variabelen herkennen en ordenen;
- verschillende diagrammen aflezen.

Verkennen

Opgave V1

Bekijk de dataset **Gegevens 154 havo 4-leerlingen** met gegevens van 154 leerlingen.

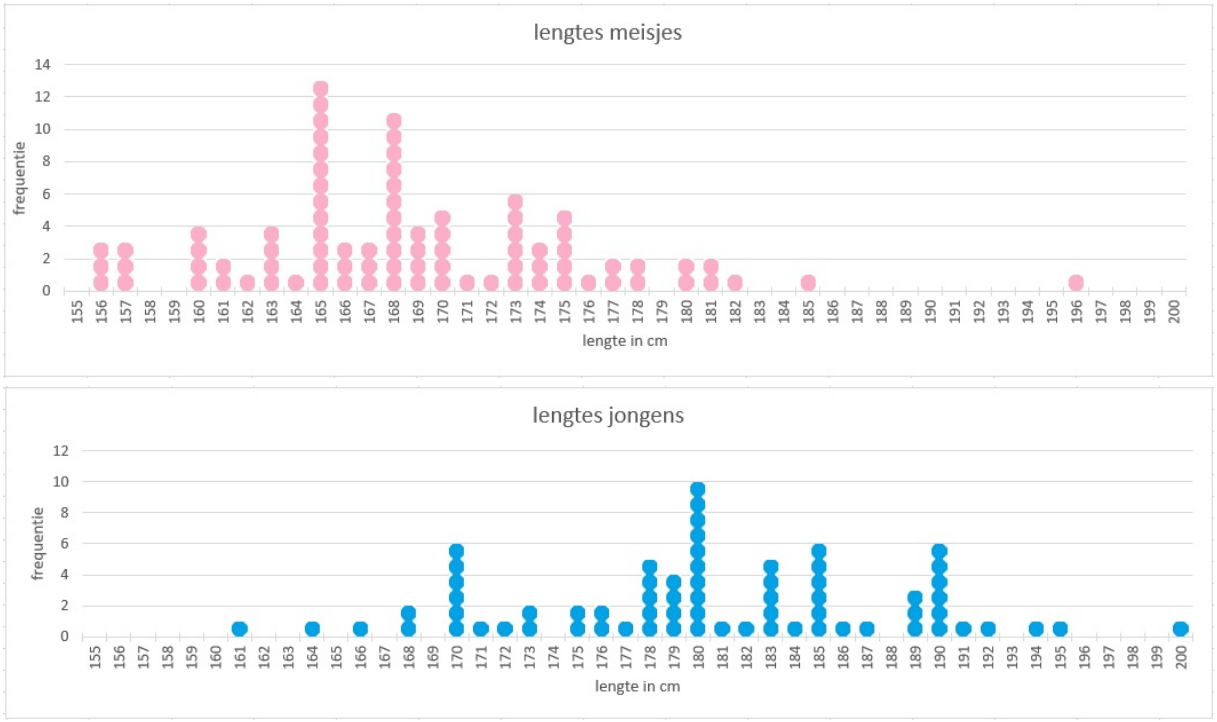
a Hoe lang is het grootste meisje?

b Hoe lang is de grootste jongen?

c Welke lengte komt het meeste voor?

d Is het berekenen van gemiddelden een goede manier om de lengtes van de meisjes en de jongens met elkaar te vergelijken? Licht je antwoord toe.

Als je de lengtes van de jongens en de meisjes wilt vergelijken, kun je ze in beeld brengen zoals hieronder.



Figuur 2

e Waarom is goed vergelijken nu nog steeds lastig?

Opgave V2

Je hebt kennisgemaakt met kwalitatieve en kwantitatieve statistische variabelen.

a Noem van beide soorten variabelen een voorbeeld.

b Aan de variabele geslacht worden soms twee waarden toegekend: 0 = vrouw en 1 = man. Wordt de variabele daarmee kwantitatief?

c De lengte bij een bevolkingsonderzoek wordt gemeten in centimeters. Kun je daarvoor redenen aangeven?

d Je ziet twee weegschalen. Wat is het verschil tussen beide als het gaat om het aflezen van een gewicht?

Figuur 3

e Bij een grafiek van het temperatuurverloop van een dag kun je een vloeiende lijn tekenen. Waarom kan dat niet bij een grafiek van de gemiddelde maandtemperatuur in een bepaald jaar?

Uitleg 1

Bekijk de dataset **Gegevens 154 havo 4-leerlingen** met gegevens van 154 leerlingen. Dit is een Excel-bestand met ruwe data. Zo'n grote brij aan gegevens is nogal onoverzichtelijk. Het werken met diagrammen helpt al iets. Hier zie je twee dotplots waarmee je de lengtes van de jongens en van de meisjes kunt vergelijken.

Bekijk de figuur in **Opgave V1**.

Zo'n dotplot is eigenlijk weinig meer dan een staafdiagram en die kun je in Excel zelf maken, zie het **Practicum**.

Omdat er 85 meisjes en 69 jongens zijn, is vergelijken nogal lastig. Het is beter om alle frequenties om te zetten naar relatieve frequenties. Zo'n relatieve frequentie wordt vaak gegeven als percentage.

Opgave 1

De dotplots in **Uitleg 1** geven de frequenties van de lengtes van de meisjes en de jongens afzonderlijk weer.

a Welke lengte komt bij de meisjes het meeste voor?

b Welke frequentie hoort daarbij?

c Hoeveel bedraagt de minimale lengte bij de meisjes?

d Hoeveel bedraagt de maximale lengte bij de meisjes?

e Bij de meisjes zit een meisje dat je met haar lengte als 'uitschieter' kunt beschouwen. Waar zit die uitschieter en waarom heet dit een uitschieter, denk je?

f Kun je op grond van wat je nu hebt gevonden de lengtes van meisjes en jongens vergelijken? En wat valt je in het bijzonder op als je de dotplots bekijkt?

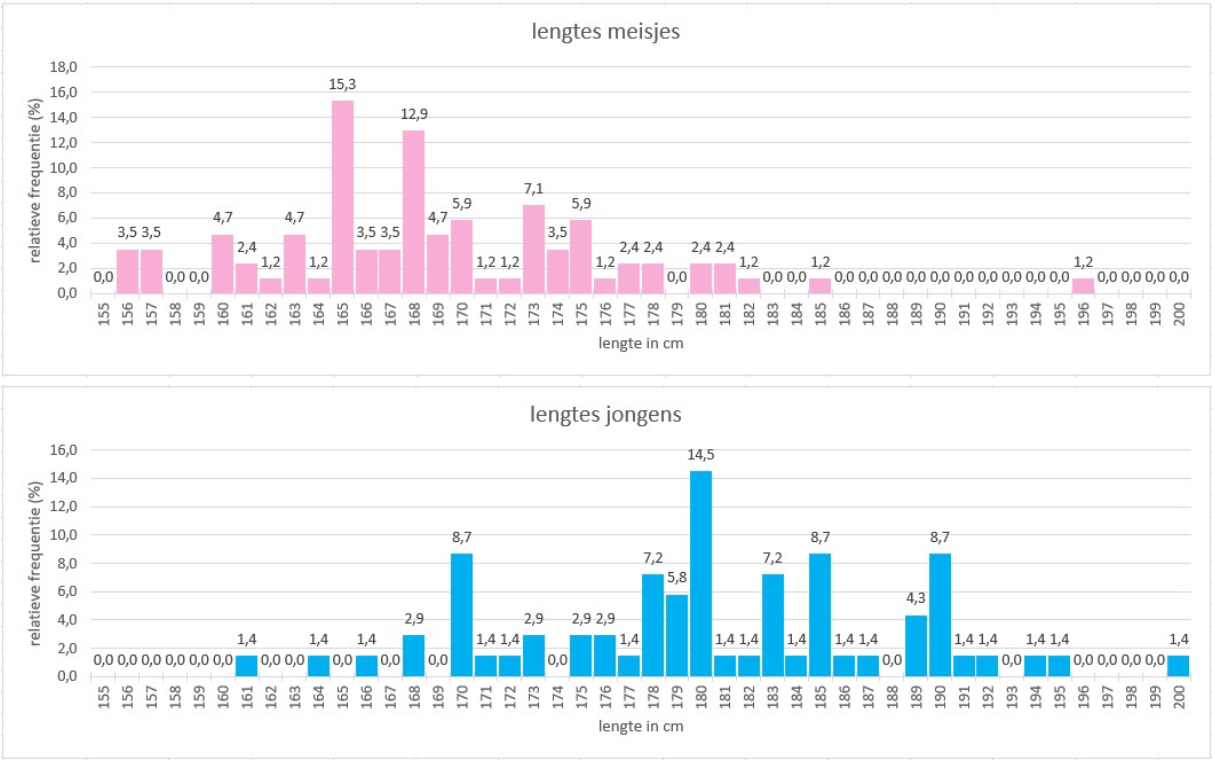
Opgave 2

Je kunt de lengtes van de jongens en de meisjes ook in staafdiagrammen zetten. Om beter te kunnen vergelijken, is het nuttig om alle frequenties uit de tabel om te zetten naar relatieve frequenties.

a Waarom is dat zo?

b Bereken het percentage dat hoort bij de lengte 175 cm bij de jongens vanuit de dotplot in één decimaal nauwkeurig. Doe dit ook voor deze lengte van de meisjes.

c Ga na of je uitkomsten per lengte overeenkomen met deze staafdiagrammen.



Figuur 4

d Hoeveel procent van de jongens is langer dan 180 cm?

e Hoeveel procent van de meisjes is langer dan 180 cm?

f Bekijk de 50% kleinste meisjes. Tussen welke waarden zit hun lengte? Hoe zit dat bij de jongens?

g Bekijk de 25% langste meisjes. Tussen welke waarden zit hun lengte? Hoe zit dat bij de jongens?

Opgave 3

Stel je maakt frequentietabellen en staafdiagrammen bij verschillende variabelen bij de 154 leerlingen uit de dataset. De volgorde waarin je de gemeten waarden zet, speelt een grote rol.

a Stel je bekijkt de variabele *profielkeuze*. Kun je daarbij een zinvol staafdiagram maken? Is de volgorde van de staven bij deze variabele van belang? En mag er tussenruimte tussen de staven zitten?

b Stel je bekijkt de variabele *huiswerk*. Kun je daarbij een zinvol staafdiagram maken?
Is de volgorde van de staven bij deze variabele van belang? Mag er tussenruimte tussen de staven zitten?

c Stel je bekijkt de variabele *geboortemaand*. Waarom is het bij deze dataset nauwelijks zinvol om bij geboortejaar een frequentietabel te maken?

d Bekijk de variabele *plezier*. Kun je daarbij een zinvol staafdiagram maken?
Is de volgorde van de staven dan van belang?
Mag er tussenruimte tussen de staven zitten?

Uitleg 2

In de dataset **Gegevens 154 havo 4-leerlingen** zitten verschillende soorten variabelen. Met kwantitatieve variabelen kun je rekenen, met kwalitatieve variabelen niet. Kwantitatieve variabelen kun je weer onderverdelen in discrete variabelen en continue variabelen.

- Discrete variabelen zijn variabelen die geen tussenwaarden kunnen aannemen. Bijvoorbeeld het geslacht, het geboortejaar, het profiel, enzovoort.
- Continue variabelen zijn variabelen als lengte, cijfergemiddelde, enzovoort. Continue variabelen kunnen allerlei tussenwaarden aannemen.

Als je dataset heel groot is, kun je ervoor kiezen om de gegevens in te delen in klassen. De daarbij horende diagrammen zijn beter met elkaar te vergelijken. In het **Practicum** kun je zien hoe dit in Excel gaat.

Deze tabel in Excel laat de lengtedata van de 154 havo 4-leerlingen in klassen verdeeld zien.

klassen			meisjes	jongens
onder	midden	boven	freq	freq
150	152,5	154	0	0
155	157,5	159	6	0
160	162,5	164	12	2
165	167,5	169	34	3
170	172,5	174	16	10
175	177,5	179	10	14
180	182,5	184	5	18
185	187,5	189	1	11
190	192,5	194	0	9
195	197,5	199	1	1
200	202,5	204	0	1
			85	69

Figuur 5

De klasse 170− < 175 is de klasse van 170 tot 175. Dat betekent dat de lengte 170 in deze klasse zit, maar dat de lengte 175 in de volgende klasse valt, namelijk in 175− < 180. Vandaar dat in Excel de bovengrens 174 is omdat alle lengtes op gehele getallen zijn afgerond. Je ziet ook de klassemiddens in de tabel. Bij een staafdiagram komen die onder het midden van elke staaf te staan.

Opgave 4

In de dataset **Gegevens 154 havo 4-leerlingen** zitten de volgende variabelen: *geslacht*, *geboortejaar*, *geboortemaand*, *gewicht*, *lengte*, *cijfergemiddelde*, *cijfer voor wiskunde*, *huiswerk*, *wiskundegroep*, *profiel* en *plezier*.
Geef voor elk van deze variabelen aan of ze kwalitatief of kwantitatief, discreet of continu zijn en welke waarden de variabelen kunnen aannemen.

Opgave 5

Om de lengtes van de 69 jongens en 85 meisjes goed te kunnen vergelijken, maak je eerst een klassenindeling en gebruik je de relatieve frequenties. In het **Practicum** kun je zien hoe dit in Excel gaat.

a Maak zelf een tabel met de relatieve frequenties bij de klassenindeling in **Uitleg 2**.

b Zet de relatieve frequenties van de lengtes van de jongens en de meisjes in aparte staafdiagrammen. Op de horizontale as komt de lengte (cm). Op de verticale as de relatieve frequentie (%).

Kun je op grond van deze staafdiagrammen bepalen hoeveel procent van de jongens langer is dan 182 cm? Licht je antwoord toe.

c Welke voordeel heeft het groeperen van de metingen in klassen? En welk nadeel?

d Welk nadeel heeft het vergroten van de breedte van de klassen?

Theorie en voorbeelden

Om te onthouden

Een verzameling gegevens van één of meer statistische variabelen is een **dataset**. Als de data niet bewerkt zijn, spreek je van **ruwe data**. Om zulke data behorende bij een variabele overzichtelijk in beeld te krijgen, maak je een **frequentieverdeling**. Zo'n frequentieverdeling heeft de vorm van een tabel, de **frequentietabel**.

Om nog beter overzicht te krijgen, zet je de gegevens uit de tabel in een diagram. Bijvoorbeeld in een **dotplot** waarin het aantal punten bij elke waarde de frequentie aangeeft, of in een **staafdiagram** waarin de lengte van de staaf de absolute of de relatieve frequentie weergeeft.

In een dataset kunnen verschillende soorten variabelen zitten, kwalitatieve of kwantitatieve variabelen. Kwantitatieve variabelen kun je nog verdelen in:

- **discrete variabelen**, deze variabelen nemen alleen vaste waarden aan.
- **continue variabelen**, deze variabelen nemen alle waarden aan.

Door de ruwe data in **klassen** in te delen, wordt je tabel overzichtelijker. Bij een klassenindeling is een gemiddelde alleen nog te schatten, omdat veel van de ruwe data in een klassenindeling niet meer terug te zien is.

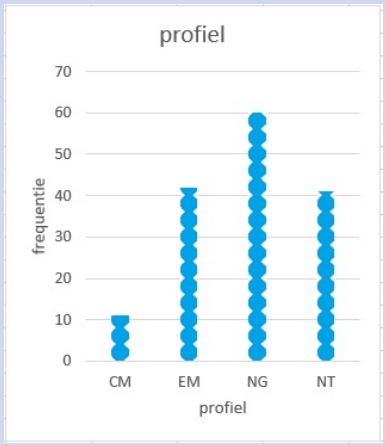
De **modus** heeft betrekking op de variabele met de hoogste frequentie. Bij klassenindelingen is de klasse met de hoogste frequenties de **modale klasse**.

Bij het indelen in klassen is er verschil tussen continue en discrete variabelen:

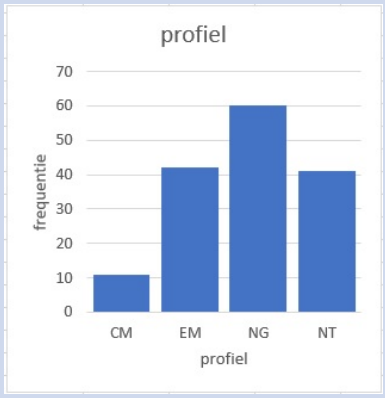
- Bij een discrete variabele, zoals *lengte afgerond op een geheel getal*, gebruik je de notatie 160 – 164 voor de klasse met alle lengtes vanaf 160 tot en met 164. De **klassenbreedte** is 5 (er zijn vijf verschillende lengtes) en het **klassenmidden** is 162.
- Bij een continue variabele, zoals *lengte*, gebruik je de notatie 160– < 165 voor de klasse met alle lengtes vanaf 160 tot aan 165. De **klassenbreedte** is 5 en het **klassenmidden** is 162,5.

Het klassenmidden gebruik je voor het schatten van het gemiddelde van een frequentieverdeling.

Bij grote datasets ontkom je er - zeker bij diagrammen - vaak niet aan om met klassenindelingen te werken.



dotplot: elke dot is 4 lln

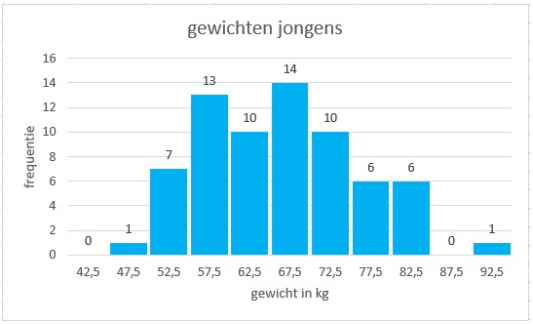


staafdiagram

Figuur 6

Voorbeeld 1

Bekijk het diagram van de frequentieverdeling van het gewicht van jongens de dataset **Gegevens 154 havo 4-leerlingen**.
Bereken bij de klasse met klassenmidden 57,5 voor deze jongens de bijbehorende relatieve frequentie. Waarom is het omrekenen naar percentages nodig als je de data van deze jongens wilt vergelijken met de data van de meisjes in deze dataset?



Figuur 7

Antwoord

Het totale aantal jongens is:
$1 + 7 + 13 + 10 + 14 + 10 + 6 + 6 + 1 = 68$.
Kennelijk heeft één van de jongens zijn gewicht niet willen geven.

Het percentage jongens bij de klasse met klassenmidden 57,5 is: $\frac{13}{68} \times 100 = 19{,}1\%$.

Omrekenen naar percentages (relatieve frequenties) is nodig om frequentieverdelingen te kunnen vergelijken. Meestal laat je Excel het rekenwerk doen.

Opgave 6

Bekijk in **Voorbeeld 1** hoe je het percentage jongens bij de klasse met klassenmidden 57,5 berekent.

a Bereken zelf het percentage jongens in de klasse $70- < 75$.

Vanuit een klassenindeling kun je het gemiddelde gewicht alleen nog maar schatten, want daarin staan niet meer de ruwe data. Je gebruikt daartoe de klassenmiddens.

b Schat het gemiddelde gewicht van deze jongens vanuit het gegeven staafdiagram.

Neem nu het databestand uit **Voorbeeld 1**.

c Bereken met behulp van Excel het gemiddelde gewicht van deze jongens.

d Maak nu zelf met behulp van Excel staafdiagrammen van de relatieve frequenties (in procenten) van de gewichten van zowel de meisjes als de jongens. Vergelijk hun gemiddelde gewichten vanuit de ruwe data.

Opgave 7

Je kunt een klassenindeling op verschillende manieren noteren.

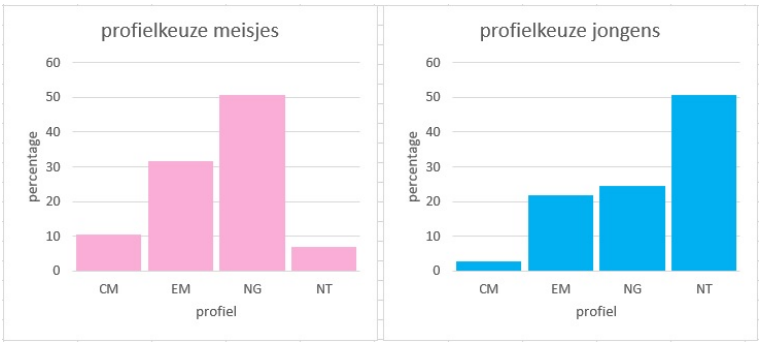
a Lengtes van de bladeren van een bepaald soort boom (cm) worden ingedeeld in de klassen $6{,}5- < 7{,}5$; $7{,}5- < 8{,}5$; enzovoort. Bepaal de klassenbreedte en de klassenmiddens.

b De leeftijden van de werknemers van een bepaald bedrijf worden in de volgende klassen ingedeeld: $20-24$, $25-29$, ..., $60-64$. Bepaal de klassenbreedte en de klassenmiddens.

c Bij een theater wordt bijgehouden hoeveel kaartjes er voor een voorstelling worden verkocht. De klasse $200-249$ geeft het aantal voorstellingen weer waarvoor 200 tot en met 249 kaartjes verkocht zijn. Bepaal de klassenbreedte en het klassenmidden van deze klasse.

d Bij welke variabelen uit een dataset is het zinvol/mogelijk om een klassenindeling te maken? Licht je antwoord toe. Gebruik de woorden kwantitatief of kwalitatief en continu of discreet.

Voorbeeld 2



Figuur 8

Bekijk de diagrammen van de frequentieverdeling van de profielkeuze van de leerlingen in de dataset **Gegevens 154 havo 4-leerlingen**.
Welke conclusies kun je trekken?

Antwoord

Je kunt beide diagrammen goed vergelijken, want in beide gevallen gaat het om relatieve frequenties.

Hieruit kun je onder andere concluderen:

- Het CM-profiel wordt weinig gekozen en meer door de meisjes dan de jongens.
- Het NG-profiel wordt vooral door de meisjes gekozen.
- Het NT-profiel wordt vooral door de jongens gekozen.

LET OP: dit geldt alleen voor deze 154 leerlingen. Je kunt nog beslist niet concluderen dat dit in het algemeen voor 4havo leerlingen geldt. Daarvoor moet je meer onderzoek doen en data verzamelen!

Opgave 8

Bekijk de twee diagrammen in **Voorbeeld 2**.

a Maak zelf de twee diagrammen.

b Hoeveel procent van de jongens kiest een N-profiel? En van de meisjes?

Je kunt ook onderzoeken hoe het zit met de keuze voor wiskunde A of B en de profielkeuze.

c Hoe zou je dat aanpakken?

Verwerken

Opgave 9

Ga van de variabelen na van welke soort ze zijn (kwalitatief of kwantitatief en discreet of continu) en welke waarden ze kunnen aannemen.

a Het *geboortejaar*.

b De *smaak van verschillende soorten rookworst*.

c De *temperatuur op de Noordpool in graden Celsius*.

d Het *gewicht van muizen in grammen*.

Opgave 10

Voor een bepaalde toets kun je maximaal 100 punten scoren. Je ziet de scores van een groep van veertig personen.

59 57 53 60 63 58 77 33 50 59 58 75 62 54 53 78 59 68 65 62
57 60 80 47 90 30 60 35 57 87 63 65 63 58 65 70 73 58 63 55

a Om welke soort variabele gaat het?

b Deel deze scores in klassen in, neem als laagste klasse $25- < 35$. Maak een frequentietabel.

c Maak bij deze tabel een staafdiagram met de relatieve frequenties.

d Personen die 55 of meer punten hebben behaald, scoren voldoende.
Hoeveel procent van deze groep scoorde voldoende?

e Je had ook als eerste klasse $30- < 40$ kunnen nemen. Wat is daarvan het nadeel?

Opgave 11

Je hebt een dataset **Sportprestaties van 74 brugklassers** (41 meisjes en 33 jongens).

In de gegevens zie je of het om een jongen of meisje gaat, de leeftijd van de leerling en de prestatie bij een sprint (in seconden), bij verspringen (cm) en bij vergooien met een kogel van 200 gram (m).

a Welke statistische variabelen tref je in deze dataset aan? Meld bij elke variabele om welke soort variabele het gaat.

b Bij het vergooien gaat het om de geworpen afstand in meters met een kogel van 200 gram. Eén jongen gooit met 40 meter het verst. Eén meisje gooit met 5 meter het minst ver. Welke klassenindeling is hierbij het meest geschikt? En als je de prestaties van de jongens en de meisjes wilt vergelijken, welke variabele kun je dan het best berekenen naar aanleiding van de frequenties? Licht je antwoord toe.

c Je ziet een frequentietabel van de vergooiprestaties met klassen van vijf meter. In Excel kun je relatief snel zelf zo'n tabel maken.

Maak met de klassenindeling staafdiagrammen voor de jongens (blauw) en de meisjes (rood) afzonderlijk in één figuur.

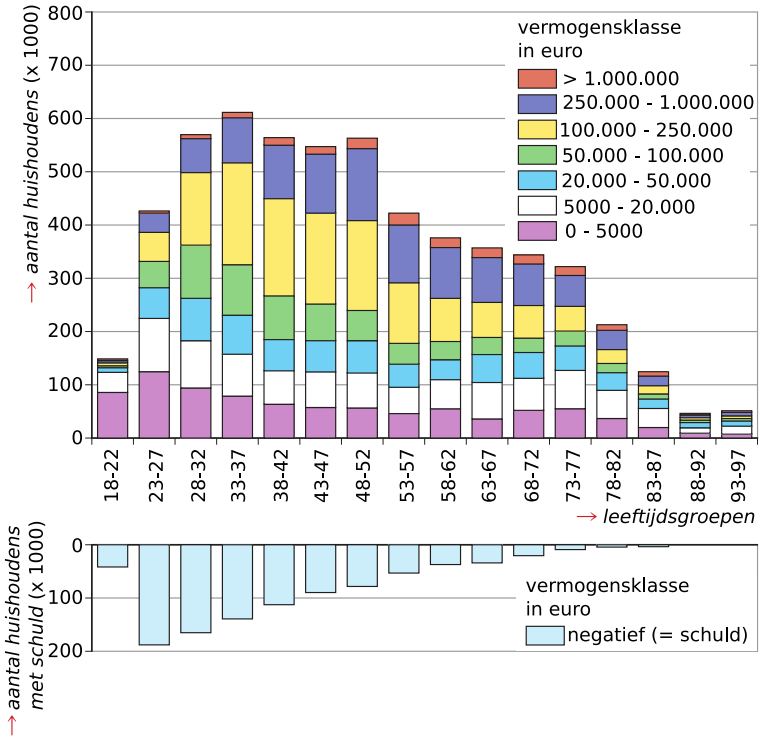
afstand			meisjes		jongens	
onder	midden	boven	freq	r.freq	freq	r.freq
0	2,5	5	0	0,0	0	0,0
5	7,5	10	1	2,4	0	0,0
10	12,5	15	11	26,8	0	0,0
15	17,5	20	18	43,9	3	9,1
20	22,5	25	9	22,0	7	21,2
25	27,5	30	2	4,9	11	33,3
30	32,5	35	0	0,0	6	18,2
35	37,5	40	0	0,0	5	15,2
40	42,5	45	0	0,0	1	3,0
			41	100,0	33	100,0

Figuur 9

d Wat valt je op bij de jongens en de meisjes wat het vergooien betreft?

Opgave 12

Het diagram geeft informatie over het vermogen of de schuld (euro) van huishoudens in land A, uitgesplitst naar de leeftijd van de hoofdkostwinner. Volgens de figuur zijn er in bijna alle leeftijdsgroepen huishoudens met een schuld.



Figuur 10

a Wat voor soort statistische variabele is *leeftijd*?

In de leeftijdsgroep 23-27 is het aantal huishoudens met een schuld het grootst.

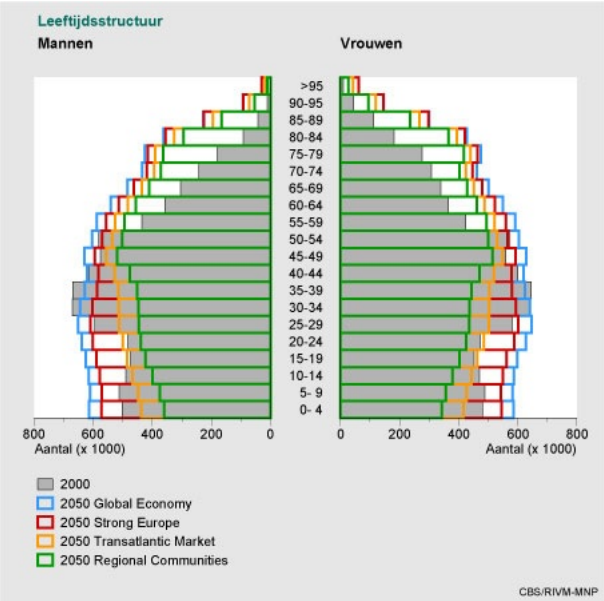
b Hoeveel procent van de huishoudens in de leeftijdsgroep van 23-27 heeft een schuld? Licht je antwoord toe en rond af op een geheel getal.

Je wilt weten hoeveel procent van de huishoudens in de leeftijdsgroep 33-37 een vermogen heeft tussen € 100000 en € 250000.

c Bereken dit percentage. Rond je antwoord af op een geheel getal.

Opgave 13

Bekijk het leeftijdsdiagram voor Nederland in het jaar 2000. Er zijn vier verschillende prognoses voor 2050 gedaan.



Figuur 11

a Bepaal de klassenbreedte en het klassenmidden van de eerste klasse.

b Hebben alle klassen dezelfde breedte?

c Waarom staan in dit leeftijdsdiagram absolute frequenties en geen relatieve frequenties?

d Je kunt dit leeftijdsdiagram omzetten naar relatieve frequenties door percentages van het totale aantal Nederlanders te nemen of door percentages van de aantallen mannen en vrouwen afzonderlijk te nemen. Noem van elke mogelijkheid een voordeel en licht je antwoord toe.

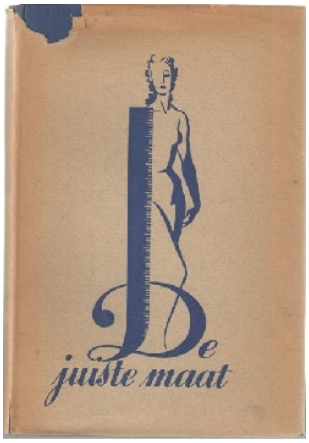
e Hoe kun je zien dat vrouwen gemiddeld langer leven dan mannen?

f De vier prognoses zijn gebaseerd op vier economische scenario's. Bij welk scenario is het vergrijzingsprobleem het sterkst in Nederland?

Toepassen

In 1951 verscheen bij uitgeverij Stafleu in Leiden het boek 'De Juiste Maat', met als ondertitel 'Lichaamsafmetingen van Nederlandse vrouwen als basis voor een nieuw maatsysteem voor damesconfectiekleding'. Auteurs van dit boek waren J. Sittig, Adviesbureau voor Toegepaste Statistiek, en prof. dr. H. Freudenthal, Rijksuniversiteit Utrecht. Het onderzoek was gehouden in opdracht van N.V. Magazijn De Bijenkorf, Amsterdam. In het kader van dit onderzoek zijn bij 5001 vrouwelijke klanten van de Bijenkorf vijftien lichaamsmaten opgemeten. Vervolgens is gekeken welke van deze maten het meest bruikbaar zijn om een **maatsysteem voor kleding** op te baseren.

Bekijk een deel van de uitkomst van het onderzoek in het bestand **Statistiek Bijenkorf 1947**.



Figuur 12

Opgave 14

Gebruik de dataset in **Toepassen**. Daarin vind je onder andere de mouwlengte van 5001 vrouwen in centimeters nauwkeurig gemeten.

mouw-lengte	frequentie
49	3
50	11
51	22
52	53
53	89
54	163
55	250
56	405
57	519
58	660
59	578
60	653
61	560
62	421
63	260
64	159
65	106
66	52
67	18
68	15
69	3
70	0
71	1

Figuur 13

a In de frequentietabel zie je de data. Met welke soort variabele heb je te maken?

b Maak een klassenindeling met klassen $45-49$, $50-54$, enzovoort. Maak daarbij een staafdiagram van relatieve frequenties in procenten nauwkeurig.

c Vergelijk deze klassenindeling met de gegeven frequentietabel en beschrijf voordelen en nadelen.

d Hoeveel procent van deze vrouwen heeft een mouwlengte van 65 cm of meer?

e Hoeveel procent van deze vrouwen heeft een mouwlengte van meer dan 65 cm?

Testen

Opgave 15

Ga van de variabelen na van welke soort ze zijn (kwalitatief of kwantitatief en discreet of continu) en welke waarden ze kunnen aannemen.

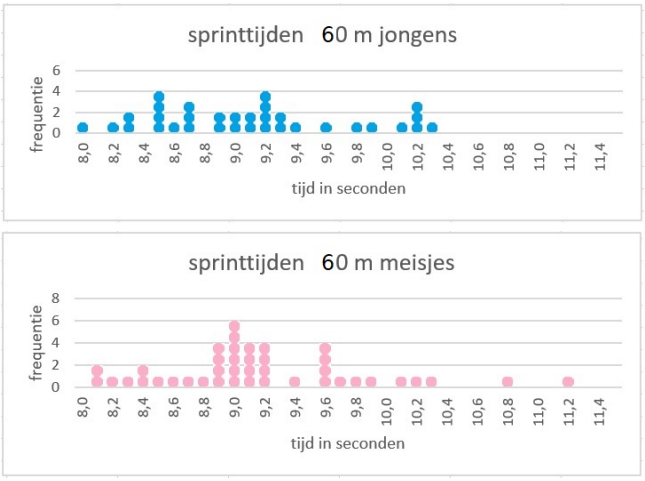
a Het *aantal dieren* van de verschillende zoogdiersoorten in een natuurgebied.

b De *hoogte* van een zoogdier.

c De *afhankelijkheid van natuurbeheer* van een diersoort.

Opgave 16

Je ziet diagrammen van de sprinttijden uit de dataset **Sportprestaties van 74 brugklassers**.



Figuur 14

a Zijn deze diagrammen dotplots?

b Welke soort variabele is hier gebruikt?

c Kun je met deze dotplots de twee deelgroepen goed vergelijken?

d Maak bij deze dotplots staafdiagrammen met relatieve frequenties en een klassenindeling van $8{,}0- < 8{,}5$, $8{,}5- < 9{,}0$, etc.

e Welke conclusies kun je nu trekken?

Practicum

Met **Excel** (een spreadsheetprogramma, een rekenblad) werken is bij statistiek eigenlijk onontbeerlijk. Je kunt er grote hoeveelheden gegevens in kwijt. Die gegevens kun je ordenen en presenteren. Bekijk de eerste drie delen van het practicum:

- **Data presenteren**

Je kunt ook data analyseren en presenteren met de app ‘Data analyse’ van **VUstat**. Daarin kun je eigen databestanden vanuit Google-Drive toevoegen, maar er zijn ook diverse datasets beschikbaar. Ga hiervoor naar:

- **Data analyse VUstat**