STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Grafieken: achtergrondinformatie

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Grafieken: achtergrondinformatie

Grafieken zijn handig om juiste informatie uit data te halen.

Een goede grafiek vertelt dikwijls veel meer dan statistische berekeningen. Het is belangrijk dat je een grafiek op de juiste manier kan interpreteren. Het is zeker ook belangrijk dat je weet hoe je een grafiek moet tekenen.

Een grafiek geeft informatie over opgemeten veranderlijken. Er zijn verschillende soorten veranderlijken. Bepaalde grafiektypes zijn geschikt om een bepaalde soort veranderlijke voor te stellen terwijl andere grafiektypes totaal ongeschikt kunnen zijn.

Naast de specifieke informatie die je wil voorstellen, bepaalt dus ook het soort veranderlijke welk grafiektype je mag gebruiken. Daarom is het goed om eerst te kijken met welke soort veranderlijke je te maken hebt.

1. Veranderlijken

Bij de meeste statistische studies trek je een steekproef uit een populatie. Voor elk **element** van de steekproef worden één of meer **veranderlijken** onderzocht. Bij een enquête kan er bijvoorbeeld gevraagd worden naar geslacht, bloedgroep, leeftijd,…
Voor de veranderlijke "geslacht" zijn er maar twee **waarden** mogelijk: mannelijk / vrouwelijk. Voor de veranderlijke "bloedgroep" heb je O, A, B, AB. De veranderlijke "leeftijd" heeft dan weer veel verschillende mogelijke waarden.

Je kan veranderlijken op verschillende manieren indelen. Bij het maken van grafieken is het handig om naar twee grote soorten te kijken:

1. veranderlijken die je als **categorisch** behandelt
2. veranderlijken die je als **continu** behandelt.

1.1. Veranderlijken die je als categorisch behandelt

Je spreekt over een **categorische** veranderlijke wanneer de mogelijke uitkomsten in (een beperkt aantal) categorieën terechtkomen. In deze inleidende tekst werken we met categorieën die elkaar niet overlappen. Elke opmeting komt terecht in één en slechts één categorie.

Uitkomsten kunnen op verschillende manieren in categorieën terechtkomen. Dat hangt af zowel van de soort veranderlijke als van de onderzoeksvraag. Hieronder zie je verschillende mogelijkheden, telkens met een voorbeeld en een opdracht.

1.1.1. Nominaal

Een **nominale** veranderlijke is categorisch. De waarden van zo'n veranderlijke kunnen alleen maar **met woorden** omschreven worden, niet met getallen.

Voorbeeld. Bij zakjes chocolade M&M-snoepjes (Choco M&M's van 45 g) kan je de kleur bestuderen. In die zakjes zitten alleen rode, groene, gele, oranje, bruine en blauwe snoepjes. De kleur is hier **de naam** van de veranderlijke. Elk snoepje komt (qua kleur) terecht in één van de 6 mogelijke categorieën: rood, groen, geel, oranje, bruin, blauw. Dat zijn de **waarden** van de veranderlijke. De **nominale** veranderlijke "kleur" is een voorbeeld van een **categorische** veranderlijke.

Opdracht 1

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen of dingen) bestudeert waarbij de opgemeten veranderlijke **nominaal** is. Geef de **naam** van de veranderlijke en haar **waarden.** Is deze veranderlijke een voorbeeld van een **categorische** veranderlijke? Waarom?

1.1.2. Ordinaal

Een **ordinale** veranderlijke is categorisch. De waarden van zo'n veranderlijke worden omschreven met woorden die een logische **orde** hebben (vandaar het woord **ordinaal**).

Voorbeeld. In een opinieonderzoek probeert men te weten te komen wat mensen denken over bepaalde onderwerpen. Hierbij gebruikt men dikwijls een enquête waarbij je één van de mogelijkheden moet aankruisen. Een voorbeeld zou kunnen zijn:

- "de huidige regering maakt haar beloften waar"
  - ☐ helemaal akkoord
  - ☐ akkoord
  - ☐ niet akkoord
  - ☐ helemaal niet akkoord

"De mate waarmee je met de uitspraak akkoord bent" is hier **de naam** van de veranderlijke. De mogelijke **waarden** van deze veranderlijke zijn: helemaal akkoord, akkoord, niet akkoord, helemaal niet akkoord. De **ordinale** veranderlijke "mate van akkoord zijn" is een voorbeeld van een **categorische** veranderlijke.

Opdracht 2

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen of dingen) bestudeert waarbij de opgemeten veranderlijke **ordinaal** is. Geef de **naam** van de veranderlijke en haar **waarden.** Is deze veranderlijke een voorbeeld van een **categorische** veranderlijke? Waarom?

1.1.3. Discreet numeriek met een klein aantal uitkomsten

Een **discreet numerieke** veranderlijke is een veranderlijke waarbij de waarden **getallen** zijn (dus **numerieke** waarden) die bovendien uit elkaar liggen (**discreet**). Als het aantal verschillende mogelijke uitkomsten "niet te groot" is, dan behandel je een discreet numerieke veranderlijke als categorisch.

Voorbeeld. Bij een studie over het aantal biologische kinderen van een vrouw noteert men het resultaat als: 0, 1, 2, 3, 4, 5, minstens 6. Het aantal biologische kinderen is hier **de naam** van de veranderlijke. Bij elke vrouw noteert men één van de 7 mogelijkheden: 0, 1, 2, 3, 4, 5, minstens 6. Dat zijn de **waarden** van de veranderlijke. De **discreet numerieke** veranderlijke "aantal biologische kinderen" kan in deze studie behandeld worden als een **categorische** veranderlijke.

Opdracht 3

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen of dingen) bestudeert waarbij de opgemeten veranderlijke **discreet numeriek** is, met een beperkt aantal verschillende uitkomsten. Geef de **naam** van de veranderlijke en haar **waarden.** Is deze veranderlijke een voorbeeld van een **categorische** veranderlijke? Waarom?

1.1.4. Continu numeriek met gegroepeerde uitkomsten

Een **continu numerieke** veranderlijke is een veranderlijke waarbij de waarden **getallen** zijn (dus **numeriek**) en waarbij de uitkomsten alle mogelijke getalwaarden kunnen aannemen tussen bepaalde grenzen, zonder enige onderbreking (**continu**). Een voorbeeld is leeftijd. Bij bepaalde onderzoeken groepeer je leeftijden in enkele categorieën, zoals: jonger dan 18, tussen 18 en 65, ouder dan 65. De **waarden** van "leeftijd" zijn in dit onderzoek dus: jonger dan 18, tussen 18 en 65, ouder dan 65. Een continu numerieke veranderlijke waarbij je de mogelijke uitkomsten groepeert in een beperkt aantal categorieën behandel je als categorisch.

Voorbeeld. De Body Mass Index (BMI) bekom je door je gewicht (in kg) te delen door het kwadraat van je lengte (in m). Deze grootheid is continu numeriek maar in veel studies kijk je alleen naar een beperkt aantal categorieën:

- BMI < 16 : ernstig ondergewicht
- $16 \leq$ BMI < 18.5 : ondergewicht
- $18.5 \leq$ BMI < 25 : normaal gewicht
- $25 \leq$ BMI < 30 : overgewicht
- BMI ≥ 30 : ernstig overgewicht (obesitas).

In deze studie is de continu numerieke veranderlijke BMI gegroepeerd in 5 categorieën die met woorden zijn omschreven. Een studie die alleen die 5 benamingen gebruikt, maakt van de continue veranderlijke een categorische (ordinaal).

Opdracht 4

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen of dingen) bestudeert waarbij de opgemeten veranderlijke **continu numeriek** is en waarbij de waarden gegroepeerd zijn in een beperkt aantal categorieën. Geef de **naam** van de veranderlijke en haar **waarden** binnen dit onderzoek**.** Is deze veranderlijke een voorbeeld van een **categorische** veranderlijke? Waarom?

1.2. Veranderlijken die je als continu behandelt

1.2.1. Continu numeriek

Een **continu numerieke** veranderlijke is een veranderlijke waarbij de waarden alle mogelijke getallen zijn tussen bepaalde grenzen. Voorbeelden zijn: gewicht, lengte, tijd, …

Als je continue uitkomsten opschrijft, dan moet je altijd ergens afronden. Het lijkt er dan op dat er tussen de verschillende waarden tussenstappen zijn, net zoals bij de discreet numerieke veranderlijke. Maar het is niet omdat je de "echte" waarde niet kan opmeten (bij gebrek aan supergevoelige meetapparatuur) of niet kan opschrijven (je moet toch ergens na de komma stoppen) dat die echte waarde er niet is. "Als model" kan zo'n "echte waarde" overal liggen.

Voorbeeld. In Vlaanderen is de gemiddelde lengte van 17-jarige meisjes 1.66 m. Bij een studie van de lengte van deze meisjes gebruik je een "min of meer" nauwkeurige meetlat en je noteert de lengte (in meter) tot op 2 decimalen. Als "model" voor de lengte van deze meisjes denk je aan een continuüm van mogelijke waarden, ergens tussen 1.20 m en 2.20 m. De naam van de veranderlijke is hier "lengte" (van 17-jarige Vlaamse meisjes) en de waarden (in m) zijn een continuüm van getallen tussen 1.20 en 2.20.

Opdracht 5

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen of dingen) bestudeert waarbij de opgemeten veranderlijke **continu numeriek** is. Geef de **naam** van de veranderlijke en haar **waarden.**

1.2.2. Discreet numeriek met een groot aantal uitkomsten

Als een **discreet numerieke** veranderlijke **een groot aantal verschillende getalwaarden** aanneemt dan kan je daarbij methoden voor een continu numerieke veranderlijke gebruiken.

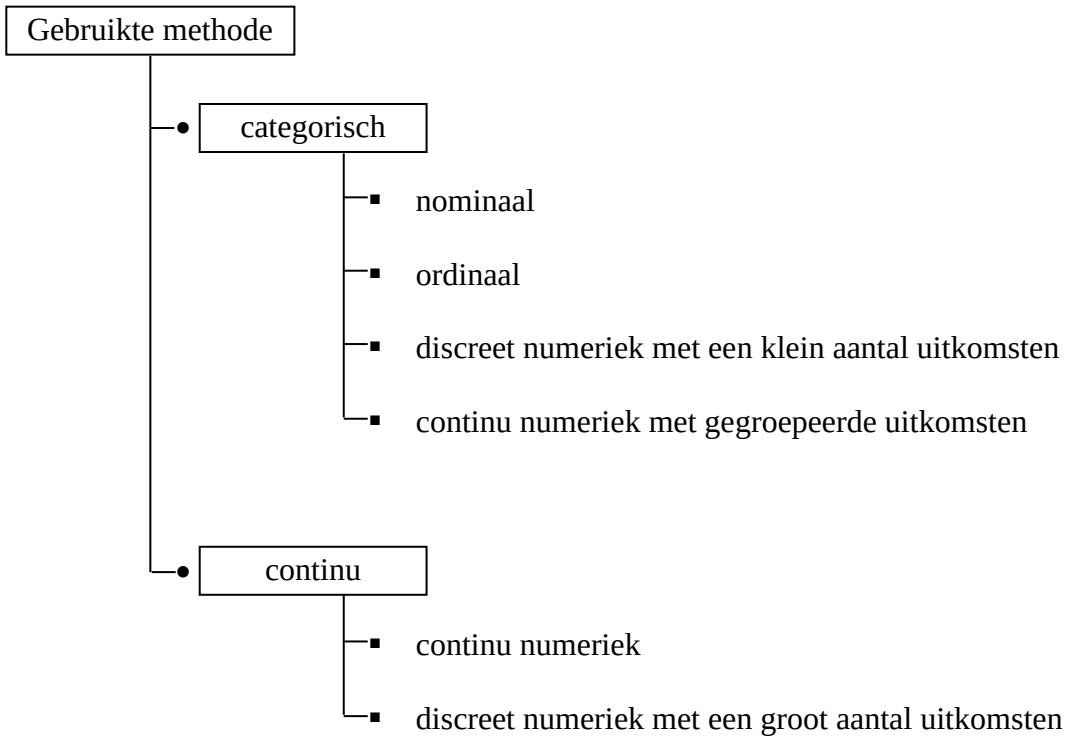
Voorbeeld. In New York zijn er duizenden hotelkamers, van zeer eenvoudige tot superluxueuze. De prijs per nacht voor zo'n kamer wordt niet door een meetapparaat opgemeten en ook niet door afronding bekomen. Het is gewoon een geheel getal (in dollar), door de hoteleigenaar vastgelegd. Als je het zo bekijkt, dan is "kamerprijs per nacht" een discreet numerieke veranderlijke. Je kan een kamer vinden van 145 dollar en misschien ook een van 146 dollar, maar je vindt er zeker geen van 145.2876 dollar voor één nacht. Je hebt hier te maken met een discreet numerieke veranderlijke met enorm veel verschillende waarden, van 15 dollar per nacht tot 15 000 dollar per nacht. Voor de studie van "kamerprijs per nacht" (= naam van de veranderlijke) kan je hier methoden voor continue veranderlijken gebruiken.

Opdracht 6

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen of dingen) bestudeert waarbij de opgemeten veranderlijke **discreet numeriek** is met **een groot aantal waarden.** Geef de **naam** van de veranderlijke en haar **waarden.**

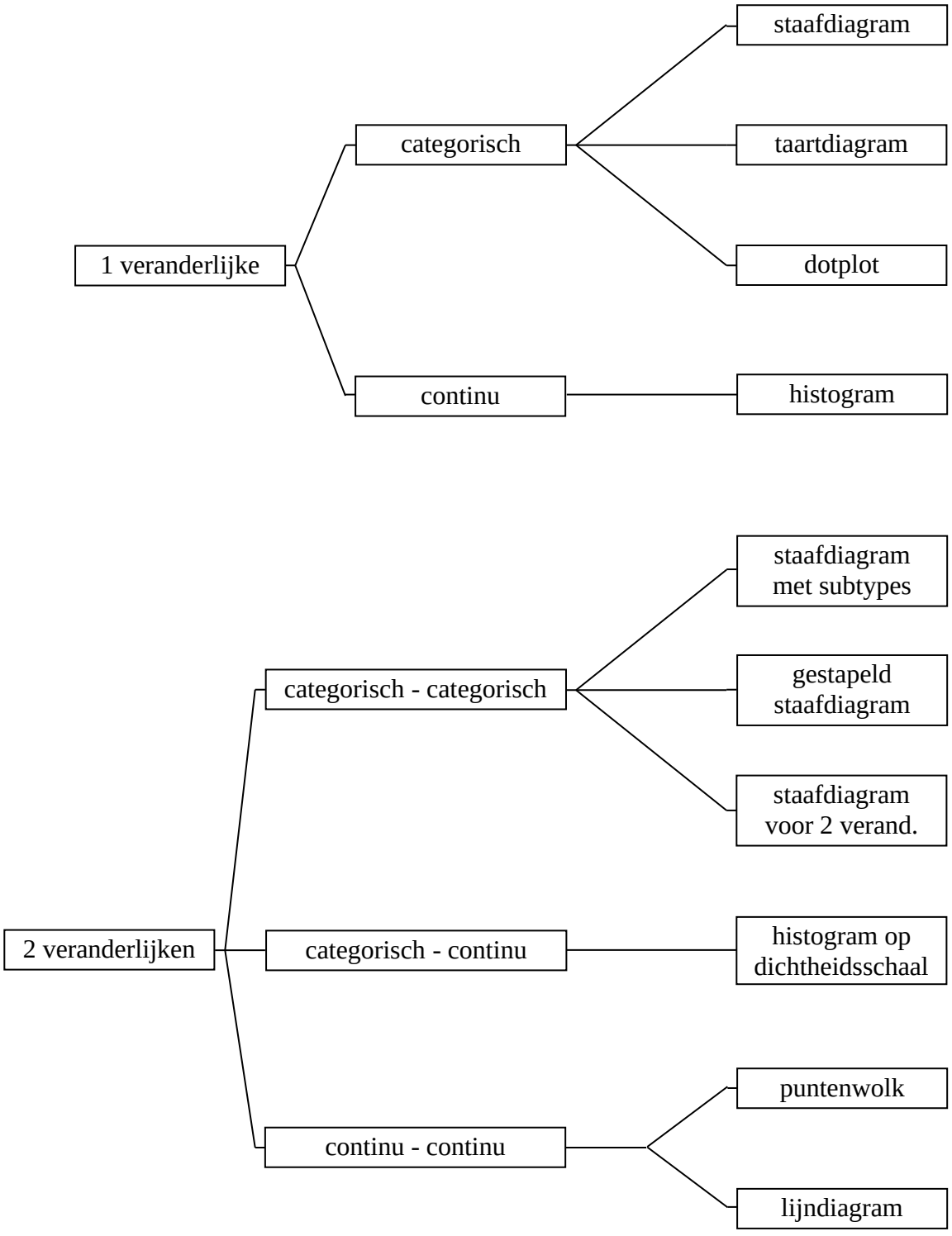
1.3. Overzicht

Bij sommige statistische studies gebruik je methoden voor categorische veranderlijken terwijl je bij andere studies kiest voor methoden voor continue veranderlijken. Onderstaand schema helpt je om de juiste keuze te maken.



2. Grafieken: overzicht

In deze tekst beperken we ons tot een klein aantal eenvoudige grafieken die veel voorkomen. Onderstaand schema geeft een overzicht.



3. Eén veranderlijke – categorisch

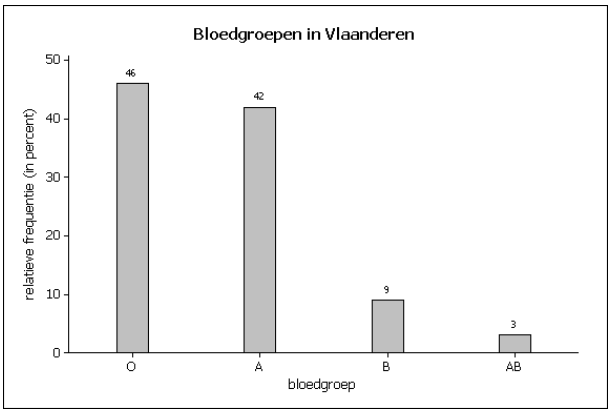
3.1. Staafdiagram

Extra informatie over het staafdiagram vind je in het lesmateriaal over *Exploratieve statistiek voor de tweede graad* en in het bijhorende *Infoboekje* op www.uhasselt.be/lesmateriaal-statistiek .

3.1.1. Staafdiagram voor een nominale veranderlijke

Voorbeeld

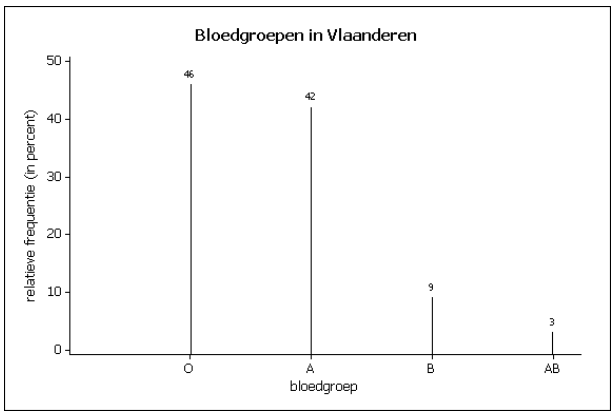
Onderstaande grafiek toont een staafdiagram voor de bloedgroepen zoals die in Vlaanderen voorkomen. De bloedgroep is een nominale veranderlijke met waarden O, A, B, AB.

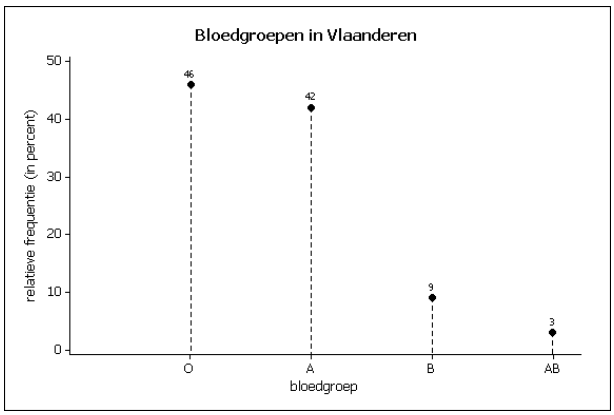


Kenmerken van de grafiek

Een staafdiagram geeft aan hoe vaak (frequentie of relatieve frequentie) een bepaalde waarde van de opgemeten veranderlijke voorkomt in het onderzoek.

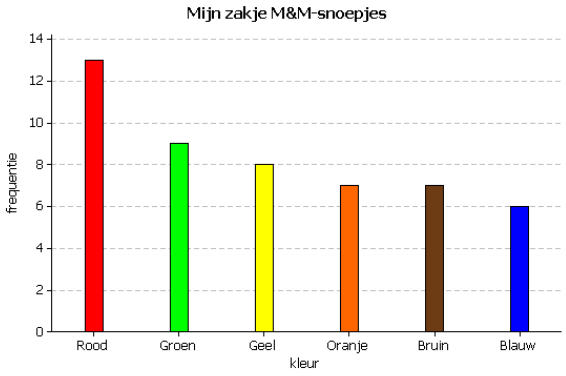
- Op de x-as staan de verschillende mogelijke waarden die de veranderlijke kan aannemen.
- Op de y-as staat de frequentie of de relatieve frequentie (als decimaal getal of in percent) van elke waarde van de veranderlijke. In dit voorbeeld zie je dat 46 % van alle Vlamingen bloedgroep O heeft.
- De waarden (op de x-as) van een nominale veranderlijke hebben geen logische volgorde. De waarden op de y-as zijn getallen en die hebben wel een volgorde. Daarom sorteer je de waarden op de x-as volgens oplopende of aflopende waarden op de y-as. Alfabetisch sorteren is bijna altijd zinloos en dat doe je dus niet.
- De waarden van de veranderlijke (de bloedgroep) komen terecht in categorieën en daarom staan deze waarden los van elkaar op de x-as. De staafjes die de frequentie (of de relatieve frequentie) vertegenwoordigen, staan bijgevolg ook los van elkaar.
- Soms tekent men staafjes, soms ook een gewoon lijntje en soms ook een punt met stippellijn (dat heet dan een dotplot). Hieronder zie je de voorbeelden.





Voorbeeld

Onderstaand staafdiagram toont hoeveel snoepjes van elke kleur er in een bepaald zakje M&M's zaten. Op de x-as staan de verschillende kleuren en op de y-as staat de frequentie (het aantal snoepjes met die kleur).

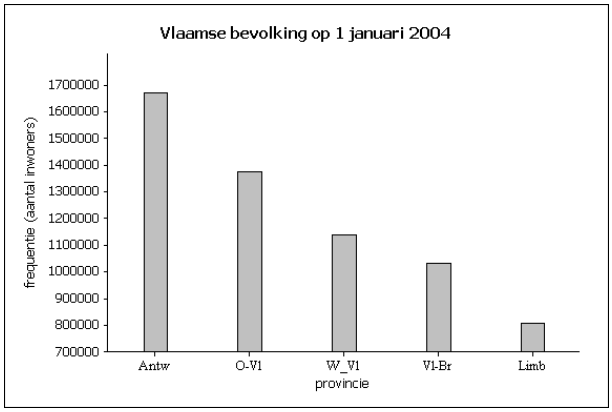


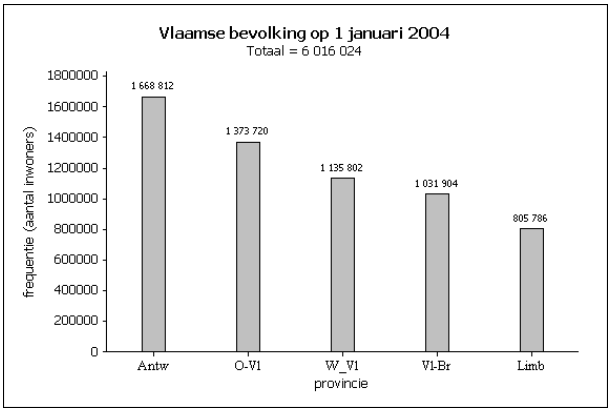
Kleur is een nominale veranderlijke en daarom is de volgorde gekozen volgens de frequenties. Wanneer twee frequenties gelijk zijn (zoals hier bij oranje en bruin), dan mag je de onderlinge volgorde vrij kiezen.

Op de figuur zie je dat er 8 gele snoepjes in dat zakje zaten.

De oorsprong van de y-as

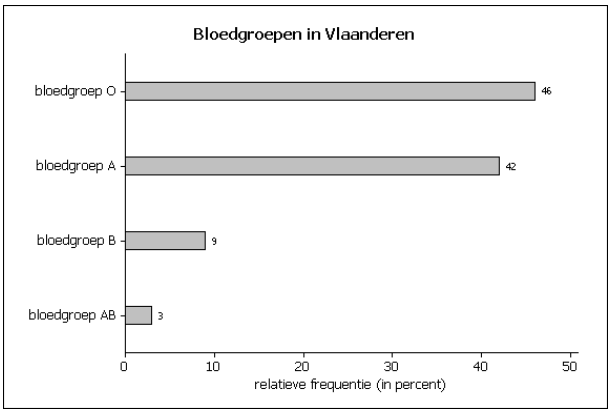
Als je naar een staafdiagram kijkt, dan vergelijk je bijna automatisch de hoogte van de staafjes met elkaar. Daarom is het nodig dat de oorsprong van de y-as zichtbaar is. Op de linkerfiguur hieronder heb je de indruk dat er in de provincie Antwerpen ongeveer 8 keer meer mensen wonen dan in Limburg. Dat is zo als je de hoogte van de staafjes met elkaar vergelijkt. De figuur zet je volledig op het verkeerde been omdat de y-as begint bij 700 000. De juiste figuur staat aan de rechterkant, waar je ziet dat er in de provincie Antwerpen twee keer zoveel mensen wonen als in Limburg.





Andere voorstellingswijze: een horizontaal staafdiagram

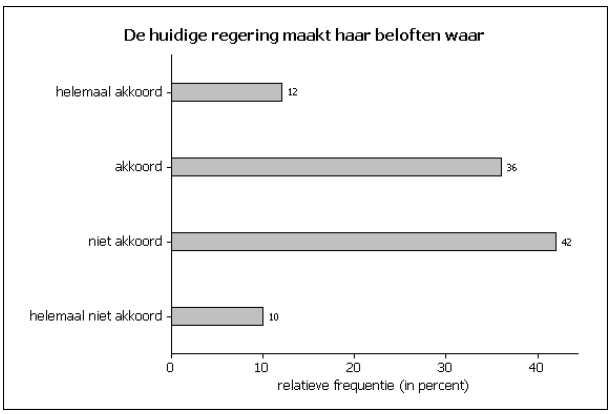
Een horizontaal staafdiagram is niets anders dan een staafdiagram waarbij de verticale as de x-as is en de horizontale as de y-as. Op die manier worden de staafjes horizontaal getekend. Dat maakt de grafiek soms beter leesbaar omdat je meer plaats hebt om de naam van de x-waarden voluit te schrijven.



3.1.2. Staafdiagram voor een ordinale veranderlijke

Voorbeeld

Onderstaand staafdiagram is een grafische voorstelling van het resultaat van een enquête waarbij gevraagd werd of je akkoord bent met de bewering dat de huidige regering haar beloften waar maakt.



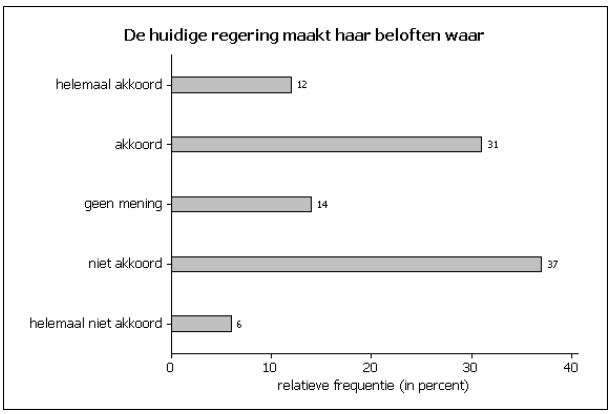
Kenmerken van de grafiek

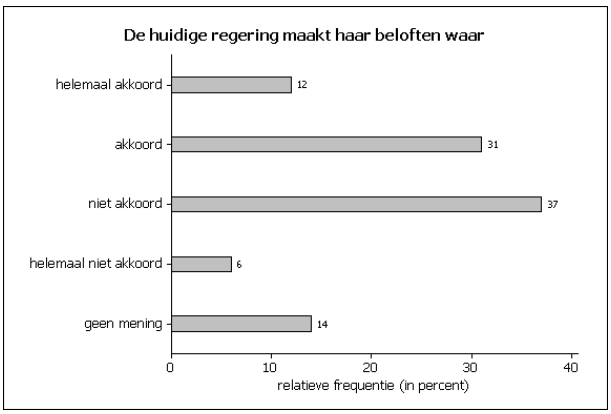
Dit staafdiagram heeft, op 1 uitzondering na, alle kenmerken van een staafdiagram voor een nominale veranderlijke. De uitzondering gaat over de soort veranderlijke. “Mate van akkoord” is een ordinale veranderlijke waarvan de waarden zelf geordend zijn. Het is dan ook logisch dat je die volgorde op de x-as respecteert.

Op de figuur zie je dat 10 % van de ondervraagden helemaal niet akkoord is met de voorgelegde uitspraak.

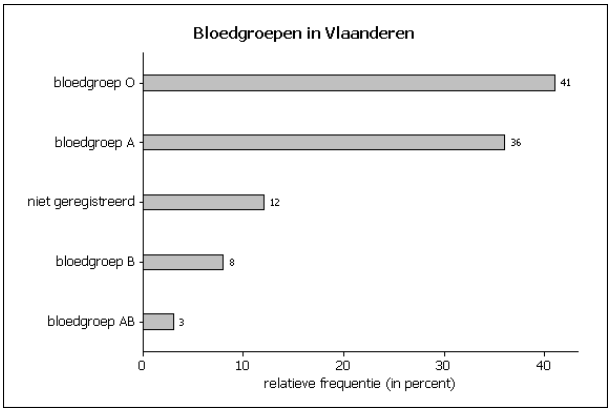
Een aparte categorie

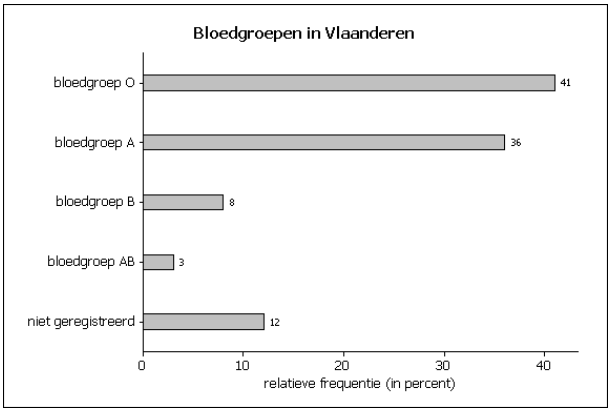
Bij heel wat enquêtes is er ook de categorie “geen mening” voorzien. Je zou dan kunnen zeggen dat die respondenten twijfelen tussen ”akkoord” en “niet akkoord” en je zou dan misschien de linkerfiguur hieronder tekenen. Dat is niet verstandig want het breekt het patroon van de mensen die wel een mening hebben. Het is beter dat je “geen mening” duidelijk apart zet, helemaal in het begin of helemaal op het einde. Daarom teken je de rechterfiguur.





Zo'n *aparte categorie* kom je regelmatig tegen, ook bij andere soorten veranderlijken. Als bij 12 % van de Vlamingen de bloedgroep niet geregistreerd is, dan teken je niet de linkerfiguur hieronder. Op die figuur geeft "niet geregistreerd" de indruk ook een bloedgroep te zijn, tussen al die andere. Op de rechterfiguur zie je dat "niet geregistreerd" iets *apart* is in deze studie. Dat trekt er de aandacht op dat je hierover iets speciaal moet zeggen bij de interpretatie van de resultaten.



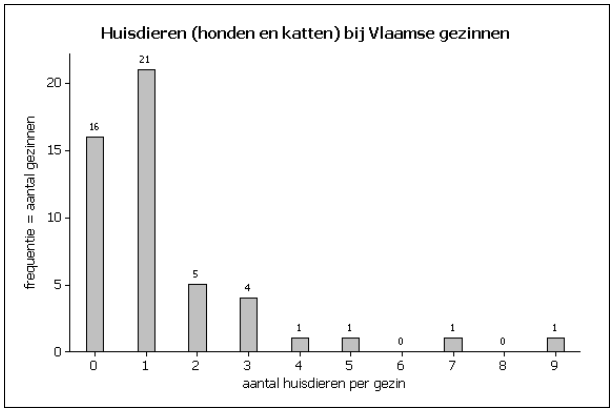


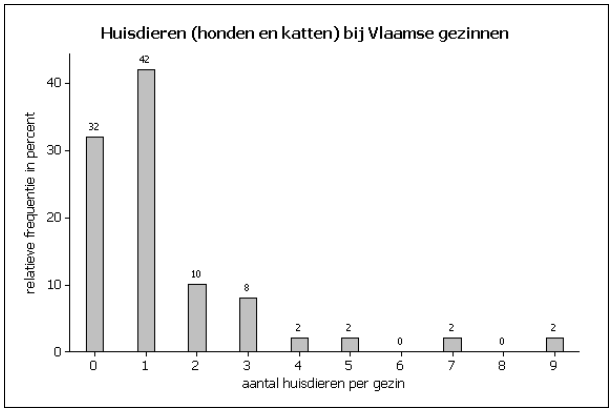
3.1.3. Staafdiagram voor een discreet numerieke veranderlijke

Een discreet numerieke veranderlijke heeft een eigen logische volgorde. Die moet je op de x-as respecteren. Ook hier kan het zijn dat je in bepaalde onderzoeken een *aparte* categorie moet afzonderen van al de rest.

Voorbeeld

Bij 50 Vlaamse gezinnen werd opgemeten hoeveel huisdieren (honden en katten) zij hadden. Het gevonden resultaat zie je in de onderstaande staafdiagrammen.





De linkerfiguur werkt met frequenties. Daar zie je dat er in die studie 16 gezinnen waren zonder huisdier. De rechterfiguur geeft relatieve frequenties in percent. Van de onderzochte gezinnen had 32 percent geen huisdier.

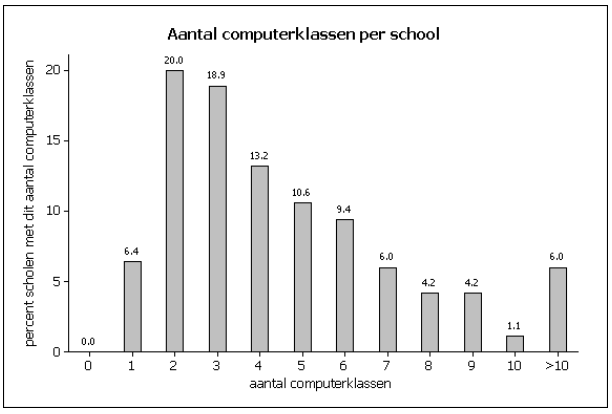
Kenmerken van de grafiek

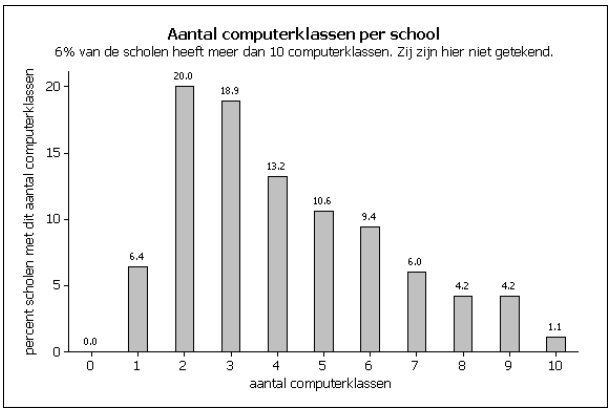
Dit staafdiagram heeft alle kenmerken van een staafdiagram voor een ordinale veranderlijke. Je moet er wel speciaal op letten dat je alle x-waarden tussen de kleinste uitkomst en de grootste op de figuur weergeeft, ook als die waarden niet in de studie voorkomen. In dit voorbeeld waren er geen gezinnen met 6 of 8 huisdieren maar die waarden staan toch op de x-as. Zo krijg je daar de juiste afstand tussen 5 en 7 en tussen 7 en 9.

Een aparte categorie

Bij sommige studies liggen bijna alle uitkomsten in een beperkt gebied (bijvoorbeeld tussen 0 en 10) behalve voor enkele gevallen die heel ver buiten dat gebied liggen (bijvoorbeeld rond 50). Als je dan een staafdiagram zou tekenen voor alle waarden, dan moet je op de x-as van 0 tot 50 gaan. Bijna alle informatie zou dan terechtkomen in een klein en weinig overzichtelijk gebied (van 0 tot 10). Daarna komt dan een grote leegte (tussen 10 en 50) en rond 50 staan dan nog enkele staafjes. Zo'n situaties kan je veel duidelijker voorstellen door "meer dan 10" als een aparte categorie te behandelen. Je kan voor die aparte categorie een staafje tekenen, of je kan ze gewoon alleen maar vermelden.

Hieronder zie je een voorbeeld. Van de onderzochte scholen heeft 20 % twee computerklassen.





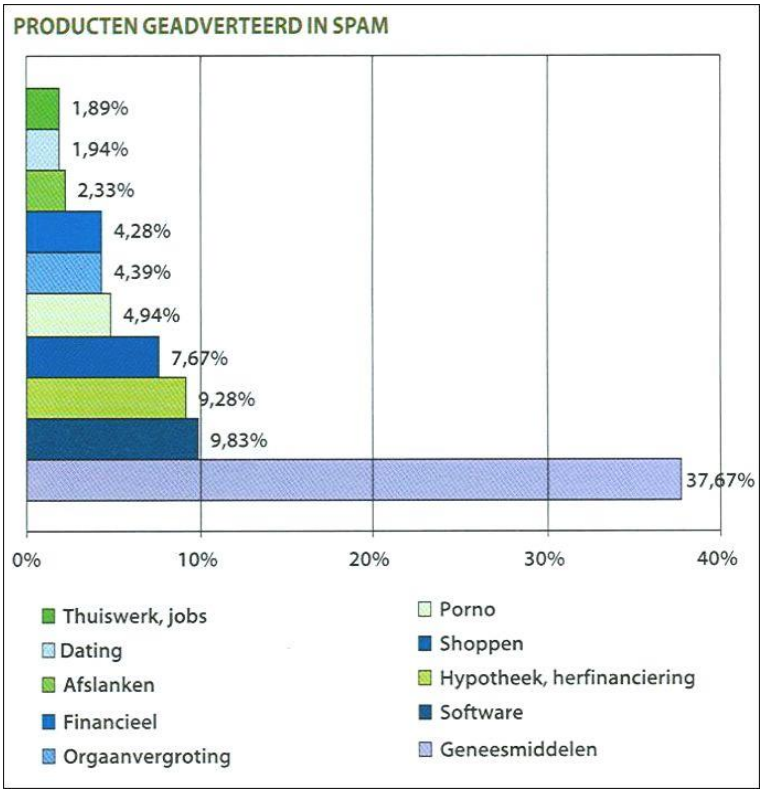
3.1.4. Opdrachten staafdiagram

VOORBEELD UIT DE MEDIA

Geneesmiddelen blijven topproduct in spam
Smart Business Strategies 1/12/2004

Korte beschrijving

Verkopers van geneesmiddelen zijn verantwoordelijk voor een steeds groter percent van de totale hoeveelheid spam: ondertussen zorgen zij al voor meer dan één derde van de mailboxvervuiling. Ook softwarehandelaars roeren zich steeds meer. Orgaanvergroting, in de eerste helft van het jaar nog op nummer 3, geraakt hierdoor helemaal in de verdrukking.



Hoe zou het onderzoek gevoerd kunnen zijn?

De elementen van de dataset zijn de spammails. Bij de onderzochte mails werden de waarden (zoals “thuiswerk, jobs”, “dating”, “afslanken”, …) van de veranderlijke (het geadverteerde product) genoteerd. Vervolgens werd voor elke waarde geteld hoeveel keer die in de studie voorkwam. Dat aantal werd dan omgezet in percent (relatieve frequentie).

Opdracht 7

Om de grafiek te verbeteren, helpt het om eerst de volgende vragen te beantwoorden.

1. Over welke soort veranderlijke gaat het in deze studie? Welke grafiek hoort daar bij? Is de afgebeelde grafiek wel van het juiste type?

2. Toont de grafiek het volledige onderzoek of maar een deel ervan? Hoe zie je dat?

3. Soms is een grafiek moeilijk leesbaar omdat je ogen telkens op en neer moeten springen bij het "decoderen" van de kleuren. Op dit punt kan je deze grafiek beter leesbaar maken.

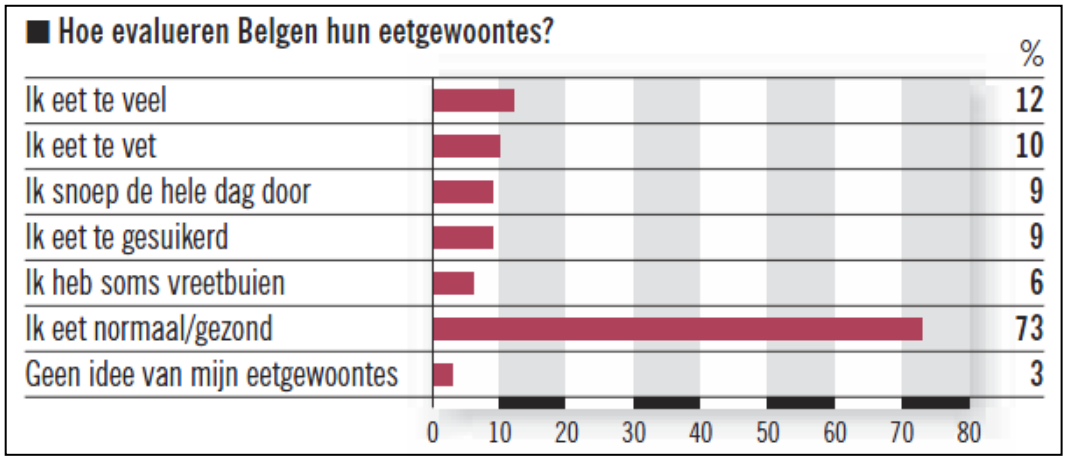
Teken nu een betere grafiek.

VOORBEELD UIT DE MEDIA

Belgen en hun gewicht
De Standaard 31/03/2004

Korte beschrijving

Het aantal Belgen dat kampt met overgewicht en zwaarlijvigheid, stijgt alarmerend snel. Uit cijfers van het Obesitas Forum blijkt dat liefst 43 procent van de Belgen een gewicht heeft dat niet echt gezond is. Het is 5 voor 12. Als we niet ingrijpen, lijdt 1 op de 10 Belgen over vijftien jaar aan suikerziekte.



Hoe zou het onderzoek gevoerd kunnen zijn?

Men heeft een steekproef getrokken uit de inwoners van België. De ondervraagde personen kregen een lijst met 7 categorieën van eetgewoontes. Zij kruisten de categorieën aan die volgens hen overeenkwamen met hun eigen eetgewoonte. Per categorie is het resultaat uitgedrukt als een relatieve frequentie in percent. Van de ondervraagden heeft 12 % “ik eet te veel” aangekruist.

Opdracht 8

Om de grafiek grondig te bespreken, beantwoord je de volgende vragen.

1. Eigenlijk zie je 3 grote groepen bij de antwoorden op dit onderzoek. Sommige mensen antwoorden “geen idee van mijn eetgewoontes” en je verwacht dat die dan alleen die categorie hebben aangekruist. Wie zegt dat hij “normaal/gezond” eet, zal ook alleen maar die categorie hebben aangekruist. De derde groep gaat over ongezond eten en daar heb je 5 mogelijkheden. Hebben alle mensen die vinden dat zij ongezond eten één enkele ongezonde categorie aangekruist? Hoe zie je dat?

2. Mag een reporter schrijven: “10 % van de Belgen vindt van zichzelf dat hij te vet eet”? Waarom?

3. Mag een reporter schrijven: “31% van de Belgen eet te veel, te vet en te gesuikerd”? Is dat in dit onderzoek correct? Waarom?

4. De getoonde grafiek is een figuur met staafjes maar het is geen staafdiagram in de gebruikelijke zin. Bij een staafdiagram teken je een verzameling staafjes die samen “het geheel” vormen. Hier zie je een grafiek waarbij de verschillende categorieën sommeren tot 122 %. Dat kan verwarring scheppen.
   De grafiek geeft informatie over de aangekruiste eetgewoonten. Je ziet bijvoorbeeld dat “te veel” dubbel zoveel keer is aangekruist als “vreetbuien”. Maar zijn die vakjes aangekruist door verschillende mensen, of door dezelfde, of wat?
   Dat je uit deze grafiek niet zomaar informatie over mensen haalt, kan je zelf mooi illustreren. Veronderstel dat je werkt met een vragenlijst bij 100 personen waarbij 3 personen antwoorden dat ze er geen idee van hebben en 73 dat ze gezond eten. Construeer voor de resterende 24 personen antwoorden die volledig overeenstemmen met de percenten in de figuur en waarbij de volgende uitspraken toch beide juist kunnen zijn:
   a. 12 % van de Belgen eet ofwel te veel, of te vet, of te gesuikerd (of een combinatie)
   b. 24 % van de Belgen eet ofwel te veel, of te vet, of te gesuikerd (of een combinatie).

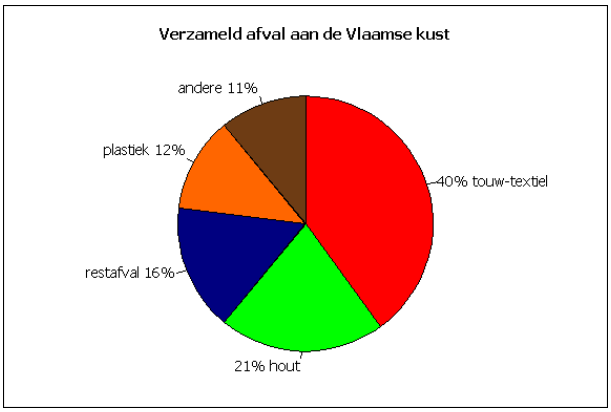
3.2. Taartdiagram

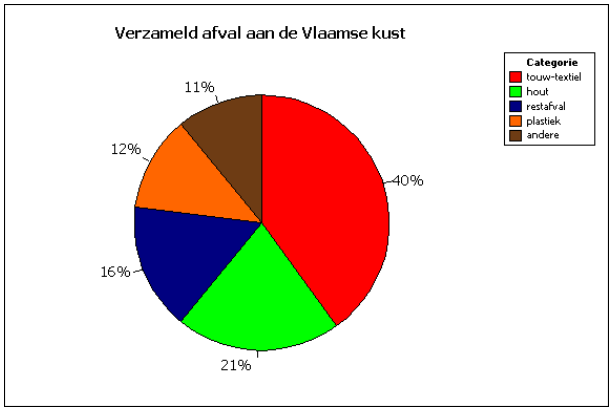
Extra informatie over het taartdiagram vind je in het lesmateriaal over *Exploratieve statistiek voor de tweede graad* op www.uhasselt.be/lesmateriaal-statistiek .

3.2.1. Taartdiagram voor een nominale veranderlijke

Voorbeeld

Hieronder zie je taartdiagrammen die een idee geven over het afval dat verzameld werd aan de Vlaamse kust tijdens een inzameldag op 1 april 2006.





Kenmerken van de grafiek

Een taartdiagram toont de relatieve frequentie (in percent) van de waarden van een categorische veranderlijk.

- De verschillende waarden van de veranderlijke moeten disjuncte categorieën vormen (er mag geen overlap zijn). In deze studie is de veranderlijke “soort afval” met waarden: “touw-textiel”, “hout”, “restafval”, “plastiek”, “andere”.
- Met elke waarde komt één sector overeen. Hier is de cirkel verdeeld in 5 sectoren.
- De volgorde waarin je de sectoren tekent, wordt meestal bepaald door de relatieve frequentie, van groot naar klein. Begin met de eerste sector op "twaalf uur" en draai dan in wijzerzin.
- Als er een categorie "andere" voorkomt, dan zijn daar de resterende waarden (metaal, glas, rubber…) in ondergebracht. Het percent van elk van die resterende waarden is kleiner dan het percent van de voorlaatste categorie, maar de som van al die resterende percenten kan soms groter zijn. De categorie “andere” plaats je helemaal achteraan, als laatste sector.
- De som van de relatieve frequenties moet gelijk zijn aan 100%.
- Bij het taartdiagram hoort een legende die duidelijk aangeeft welke waarde bij welke sector hoort. Soms staat de legende in een afzonderlijk kader en dan moeten je ogen heen en weer bewegen om te weten welke sector bij welke waarde hoort. Dat zie je op de rechterfiguur. Meestal is een taartdiagram beter leesbaar als je de legende bij de sectoren schrijft, zoals op de linkerfiguur.

Aandachtspunten

- Een taartdiagram toont één veranderlijke. Die veranderlijke moet categorisch zijn met categorieën die niet overlappen want je verdeelt de taart in een aantal niet-overlappende sectoren. Hoeveel graden een sector is bekom je uit: (relatieve frequentie) x (360°).

- Een sector van nul graden zie je niet en die teken je ook niet. Een taartdiagram is geen aangewezen figuur voor een onderzoek waarbij sommige waarden ontbreken (zoals bij het aantal huisdieren bij Vlaamse gezinnen).

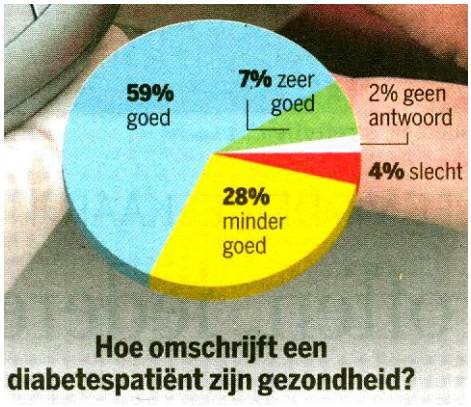
3.2.2. Taartdiagram voor een ordinale veranderlijke

Als je een taartdiagram ontmoet dan gaat het meestal over een nominale veranderlijke. Soms zie je ook een taartdiagram voor een ordinale veranderlijke. Zo'n veranderlijke heeft zelf een logische volgorde. Die volgorde respecteer je, juist zoals bij een staafdiagram voor een ordinale veranderlijke.

VOORBEELD UIT DE MEDIA

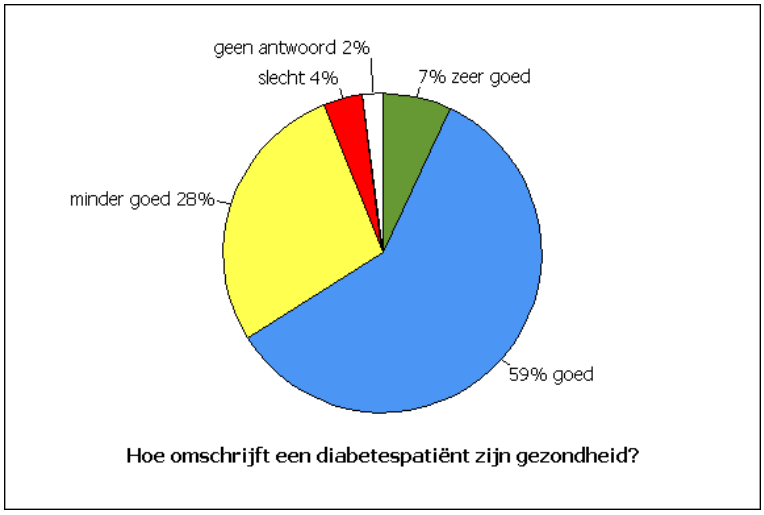
Diabetes neemt sterk toe
De Standaard 15/03/2012

Korte beschrijving



Bij diabetes produceert het lichaam onvoldoende insuline of is het insuline–ongevoelig. Daardoor ontstaat een verhoging van het bloedsuikergehalte. Diabetes wordt soms ook suikerziekte genoemd. Het is een ongeneeslijke ziekte.
De verstoring van het insulinepeil kan liggen aan een foute werking van de pancreas ("type 1 diabetes") of aan een slechte werking van de cellen die de suikers moeten opnemen ("type 2 diabetes"). Vooral type 2 komt almaar vaker voor, deels door de veroudering van de bevolking, deels door slechte voeding, overgewicht en gebrek aan beweging.
De cijfers zijn afkomstig van een onderzoek bij leden van het socialistisch ziekenfonds.

Kenmerken van de grafiek



Een taartdiagram voor een ordinale veranderlijke is zoals een taartdiagram voor een nominale veranderlijke waarbij de volgorde van de sectoren nu bepaald wordt door de logische volgorde van de ordinale veranderlijke. Die eigenschap is in de krant gerespecteerd. De figuur kan nog iets duidelijker door 3D weg te laten, te starten op "twaalf uur" en te draaien in wijzerzin.

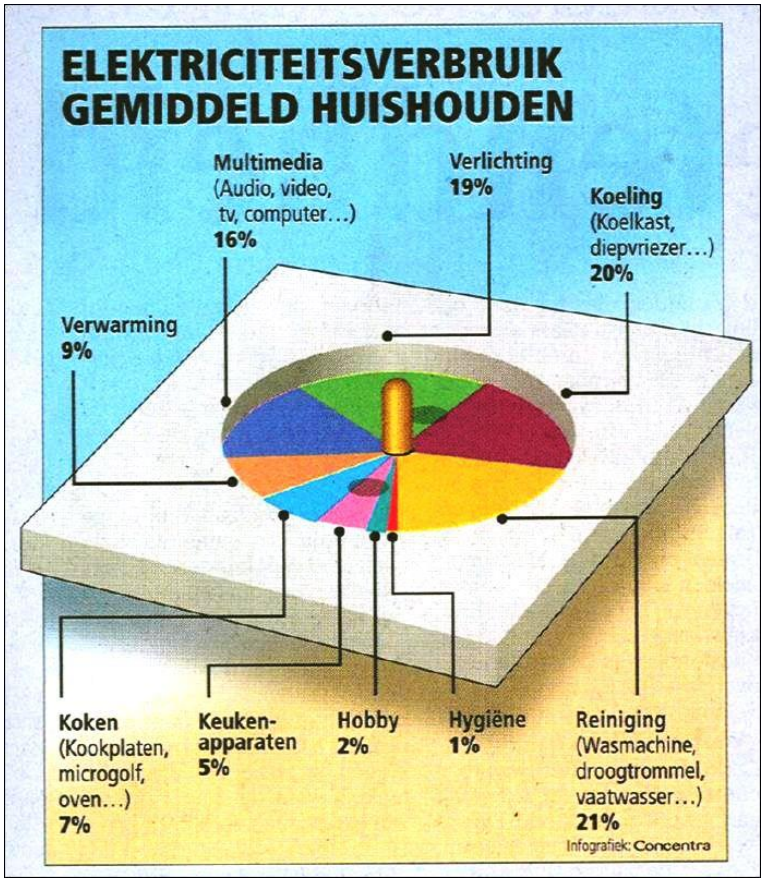
3.2.3. Opdrachten taartdiagram

VOORBEELD UIT DE MEDIA

Elektriciteitsverbruik gemiddeld huishouden
Het belang van Limburg 22/10/2004

Korte beschrijving

Wie dacht dat de torenhoge olieprijzen enkel vat hebben op de diesel-, benzine-, stookolie- en gasprijzen, zou wel eens bedrogen kunnen uitkomen. Ook de prijs van onze propere elektriciteit is immers (gedeeltelijk) afhankelijk van de olieprijs. Hoe is het elektriciteitsverbruik verdeeld in een gemiddeld huishouden?



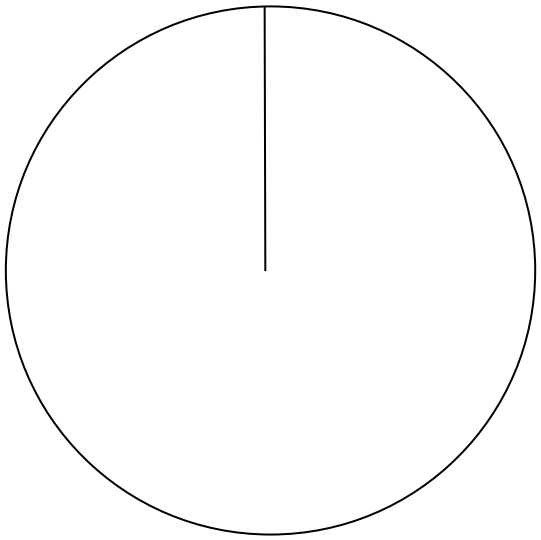
Hoe zou het onderzoek gevoerd kunnen zijn?

Bij een aantal huishoudens is "de verdeling van het elektriciteitsverbruik" opgemeten. Dit is de bestudeerde veranderlijke, met waarden: "voor reiniging", "voor koeling", "voor verlichting",… Hier is gewerkt met 9 categorieën die elkaar niet overlappen. Voor elke categorie is het gemiddelde berekend. Hoe de steekproef van de huishoudens is getrokken staat niet vermeld.

Opdracht 9

1. De getoonde grafiek voldoet aan verschillende kenmerken van een taartdiagram. De grootste sector start niet bij "twaalf uur" maar dat is geen fout. Starten bij twaalf uur is enkel een conventie die handig is wanneer je zelf een taartdiagram moet tekenen.
   Noem 3 kenmerken van een goed taartdiagram waaraan deze grafiek voldoet.

2. De grafiek gebruikt zowel een perspectiefweergave (gekantelde taart) als 3D. Dat bemoeilijkt een goede eerste indruk. Zo lijken de blauwe en paarse sectoren even groot, maar dat is niet zo. Een taartdiagram is duidelijker als je gewoon vlak in 2D tekent. Maak nu zo'n verbeterde figuur.

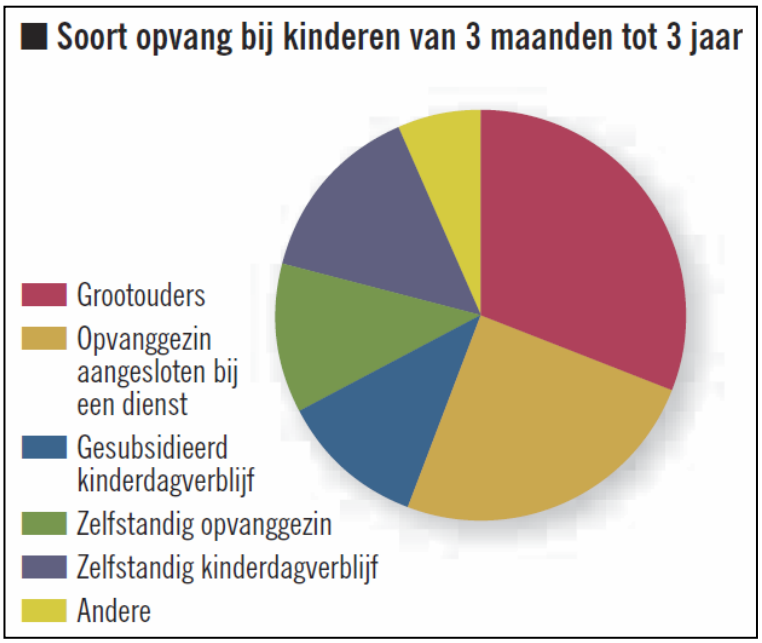


VOORBEELD UIT DE MEDIA

Almaar meer kinderen in de opvang
De Standaard 29/06/2004

Korte beschrijving

Vijftigduizend extra plaatsen in de kinderopvang tot drie jaar, en tienduizend in de buitenschoolse opvang. Dat is voor Kind en Gezin de minimale inzet tegen het jaar 2010. "Investeren in de vroege kinderjaren rendeert nog meer dan in het onderwijs".



Hoe zou het onderzoek gevoerd kunnen zijn?

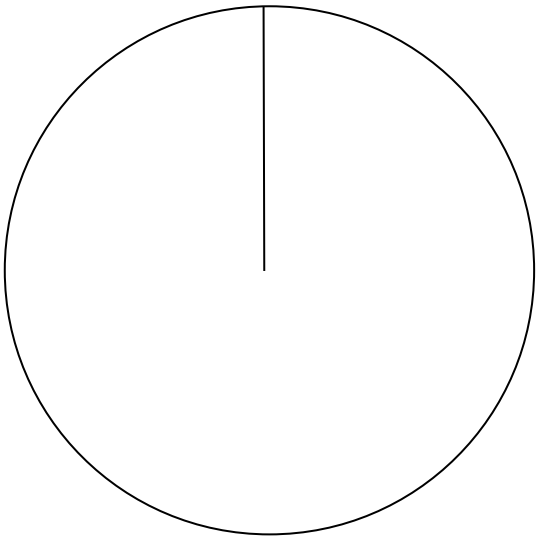
De soort kinderopvang (= de veranderlijke) is onderzocht bij kinderen van 3 maand tot 3 jaar die naar een kinderopvang gingen. Daarbij is er een onderverdeling gemaakt in 6 categorieën. Als je veronderstelt dat elk kind in slechts één categorie is ingedeeld, dan overlappen de categorieën elkaar niet. Met elke categorie kan je dan een cirkelsector laten overeenstemmen. De categorie "andere" wijst erop dat niemand vergeten is zodat de verschillende sectoren sommeren tot 100%.

Opdracht 10

1. De getoonde grafiek mist duidelijkheid. De grootste sector begint bij “twaalf uur” en dan volgt de tweede grootste sector. Dat is goed. Daarna echter, bij de groene en blauwe sectoren, krijg je de indruk dat de volgorde niet helemaal klopt. Je kan dat ook niet controleren want er staan geen percenten bij. Wat denk je van de sectoren “blauw – groen – grijs”? Hoe zou je die ordenen qua grootte?

2. Teken een verbeterde grafiek die goed leesbaar is en waarbij je gebruik maakt van de volgende gegevens.

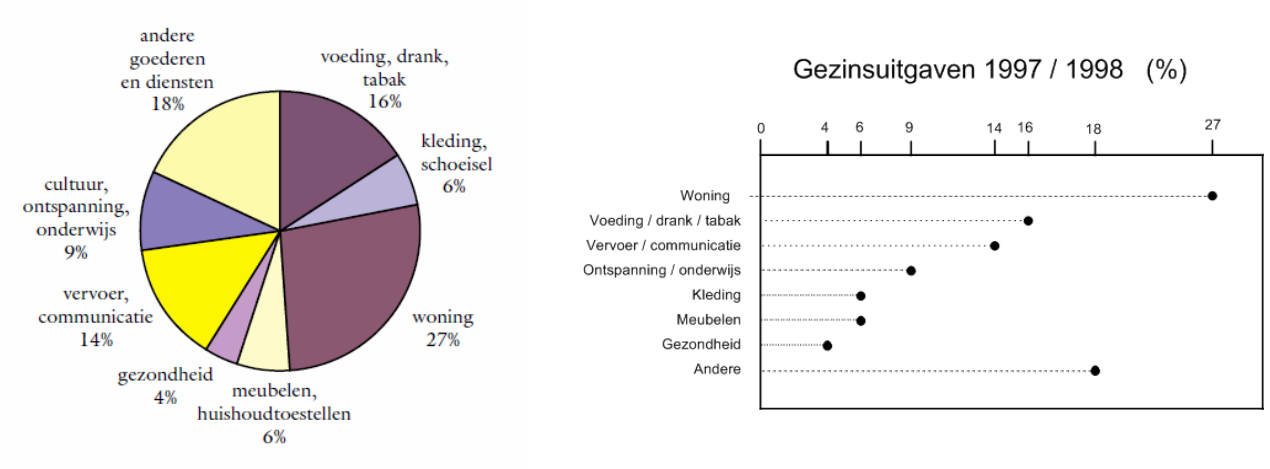
Grootouders	31 %
Opvanggezin aangesloten bij een dienst	25 %
Gesubsidieerd kinderdagverblijf	12 %
Zelfstandig opvanggezin	12 %
Zelfstandig kinderdagverblijf	14 %
Andere	6 %



3.3. Dotplot

Een dotplot is een speciaal geval van een staafdiagram. Een dotplot gebruik je hoofdzakelijk bij nominale veranderlijken. In vergelijking met een gewoon staafdiagram of een taartdiagram is een dotplot dikwijls efficiënter om patronen te ontdekken.

In de publicatie "*Profiel Vlaanderen 2000*" van het Ministerie van de Vlaamse Gemeenschap staan de gezinsuitgaven van 1997/1998 geïllustreerd met een taartdiagram. Als je deze gegevens omzet in een dotplot dan zie je bepaalde clusters veel beter.



Een dotplot kan handig zijn als je een nominale veranderlijke bestudeert die redelijk veel waarden aanneemt. We illustreren dit met een voorstelling van het aantal inwoners in West-Vlaanderen.

Hieronder zie je 4 grafieken. Zij illustreren niet alleen het gebruik van een dotplot maar zij laten je ook zien hoe je een grafiek beetje bij beetje kan verbeteren.

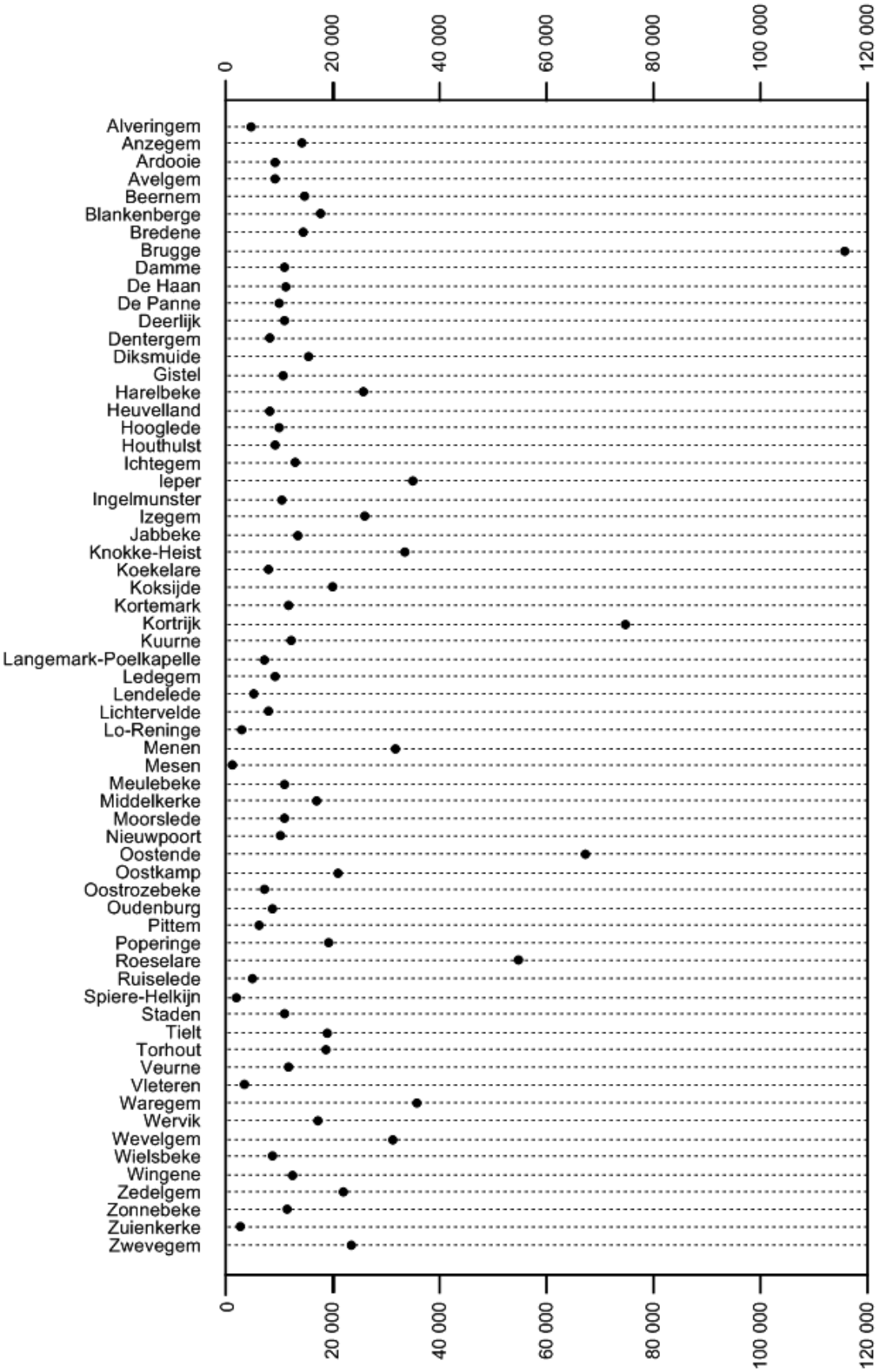
1. De eerste grafiek zou je waarschijnlijk spontaan maken: je plaatst de gemeenten in alfabetische volgorde. Zo zie je snel waar Kuurne staat, maar het is moeilijk om te ontdekken welke gemeenten ongeveer evenveel inwoners hebben als Kuurne. Er zit geen patroon in.

2. De tweede grafiek is geordend volgens aantal inwoners. Hier zie je onmiddellijk de grote clusters: Brugge is een buitenbeentje, dan heb je de groep Kortrijk-Oostende-Roeselare, enz.

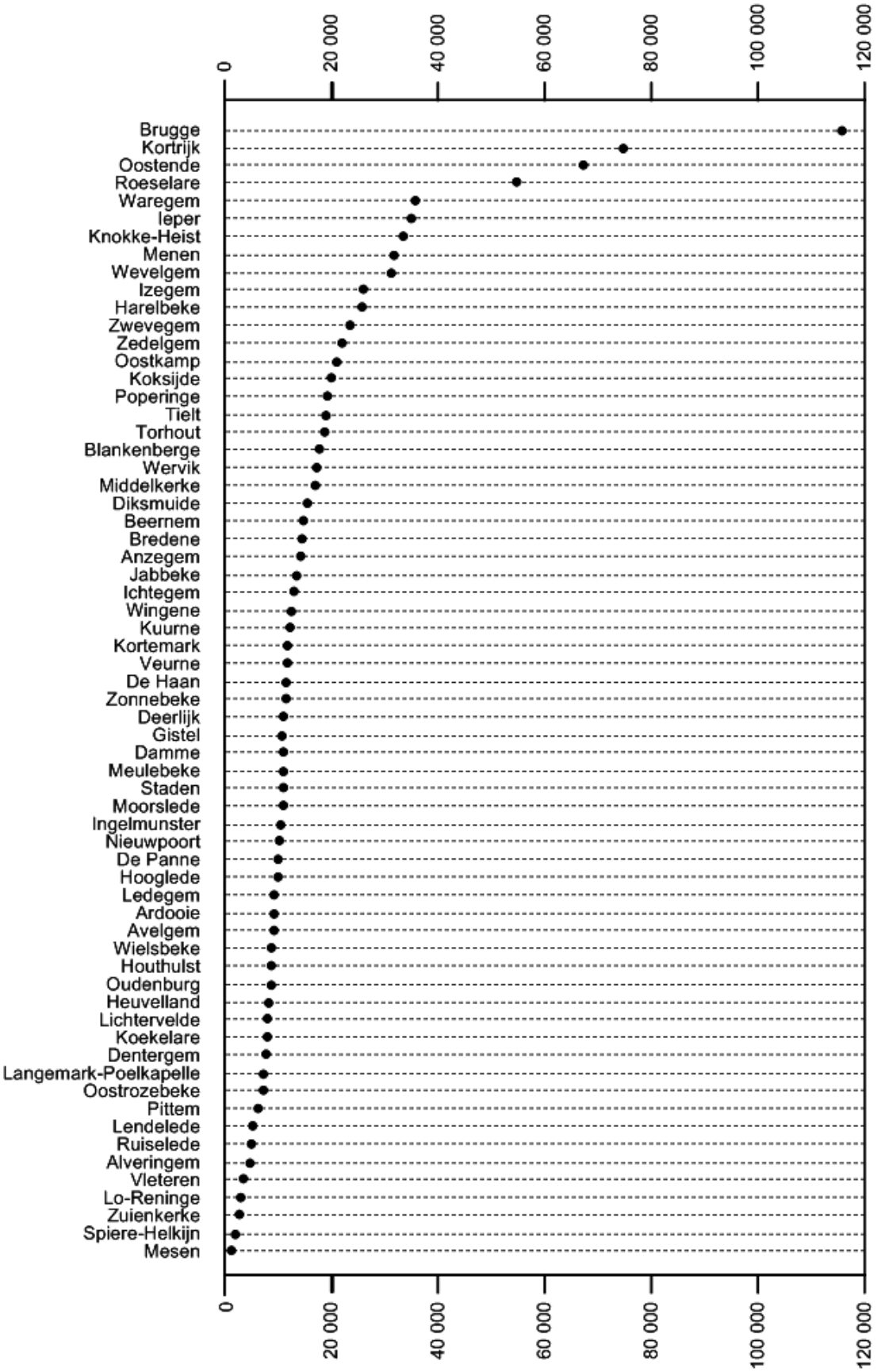
3. De derde grafiek is zoals de tweede, maar met extra hulp voor je ogen om gemakkelijker de grootteorde te schatten. Hier zie je beter dat Kuurne rond de 12 000 inwoners heeft. Op de tweede grafiek is dat veel moeilijker te zien. Je ontdekt hier ook dat de "buren van Kuurne" Wingene en Kortemark zijn. Dat kan je op de eerste grafiek bijna niet zien.

4. Op de tweede en derde grafiek is het moeilijker dan op de eerste om snel te vinden waar Kuurne staat. De vierde grafiek helpt je met een "referentietabel". Je kan snel iets ontdekken in alfabetische volgorde en je kan dat ook in numerieke volgorde. In de referentietabel vind je snel Kuurne (alfabetisch) en daarnaast staat het nummer 29. Dat nummer vind je ook snel als je in de grafiek van boven naar beneden loopt.

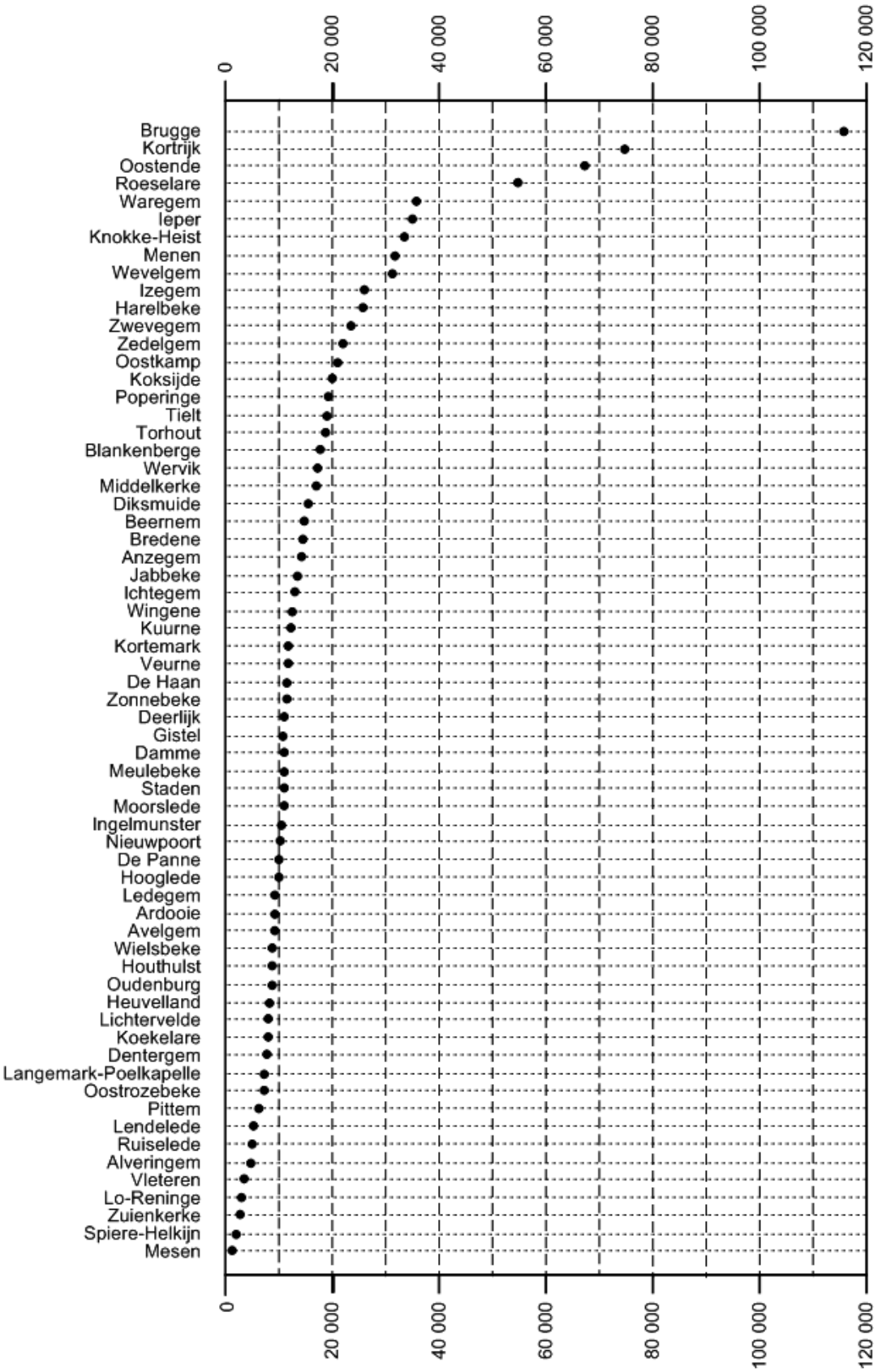
Provincie West-Vlaanderen. Bevolking per gemeente (2002)



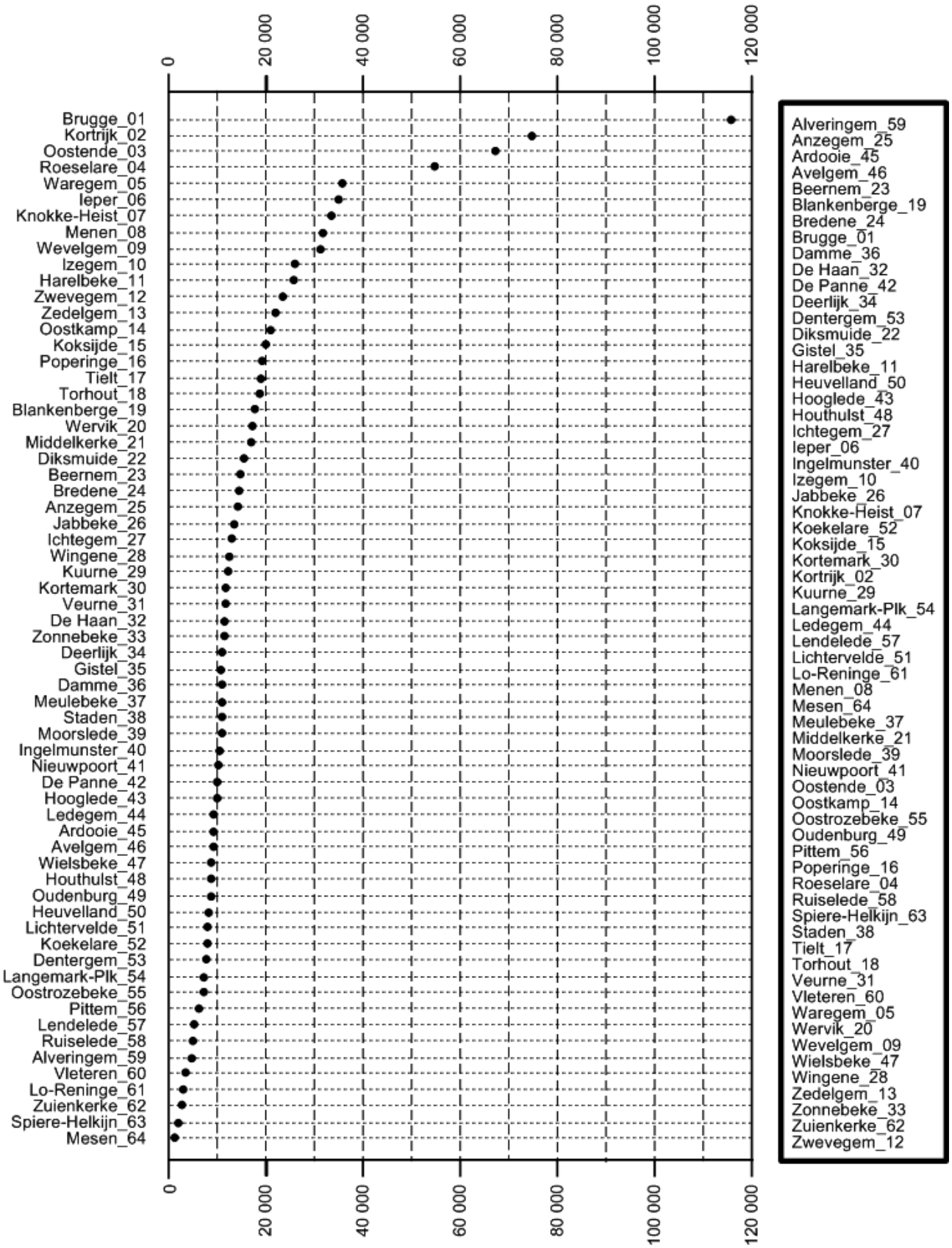
Provincie West-Vlaanderen. Bevolking per gemeente (2002)



Provincie West-Vlaanderen. Bevolking per gemeente (2002)



Provincie West-Vlaanderen. Bevolking per gemeente (2002)



4. Eén veranderlijke – continu

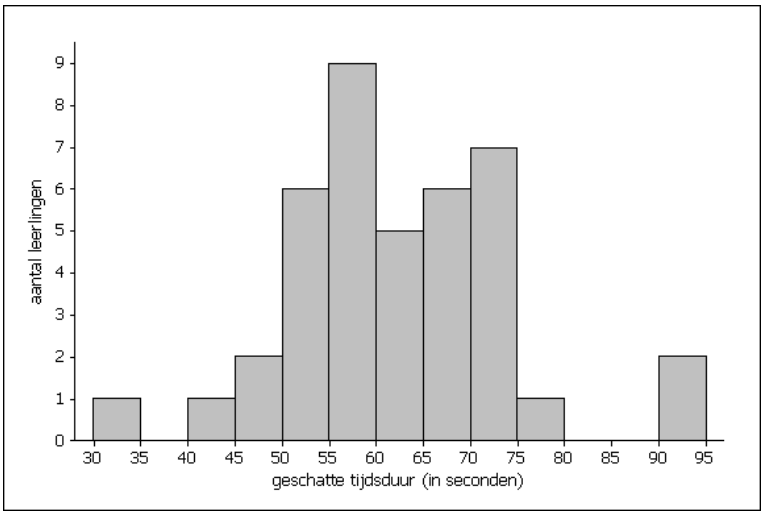
4.1. Histogram

Extra informatie over het histogram vind je in het lesmateriaal over *Exploratieve statistiek voor de tweede graad* en in het bijhorende *Infoboekje* op www.uhasselt.be/lesmateriaal-statistiek .

4.1.1. Histogram voor een continu numerieke veranderlijke

Voorbeeld

In onderstaand histogram zijn de resultaten voorgesteld van een onderzoek dat is uitgevoerd in een school te Diest. Via een steekproef werden 40 leerlingen geselecteerd die met gesloten ogen de tijdsduur van één minuut moesten schatten. De geschatte tijdsduur werd genoteerd tot op de seconde.



Kenmerken van de grafiek

Een histogram is de meest gebruikte figuur om het globale gedrag van een continue veranderlijke voor te stellen. In deze studie gaat het over “geschatte tijdsduur”. Als model is dat een continue veranderlijke, ook al zijn de opmetingen afgerond tot op de seconde.

Een histogram teken je als volgt:

- De waarden van de veranderlijke worden verdeeld in aaneengesloten klassen, die aangeduid worden op de x-as. Je kan de klassengrenzen aangeven of, als alle klassen even breed zijn, enkel het klassenmidden.
- Op elk interval komt een rechthoek. Aangezien alle intervallen op elkaar aansluiten, liggen ook alle rechthoeken tegen elkaar.
- De hoogte van de rechthoek mag je zelf kiezen, zolang je de basiseigenschap van een histogram respecteert. Die basiseigenschap zegt dat **de oppervlakte** recht evenredig moet zijn met het aantal observaties in de klasse. Als je een histogram op de dichtheidsschaal tekent, dan is de oppervlakte gelijk aan het percent observaties in de klasse.

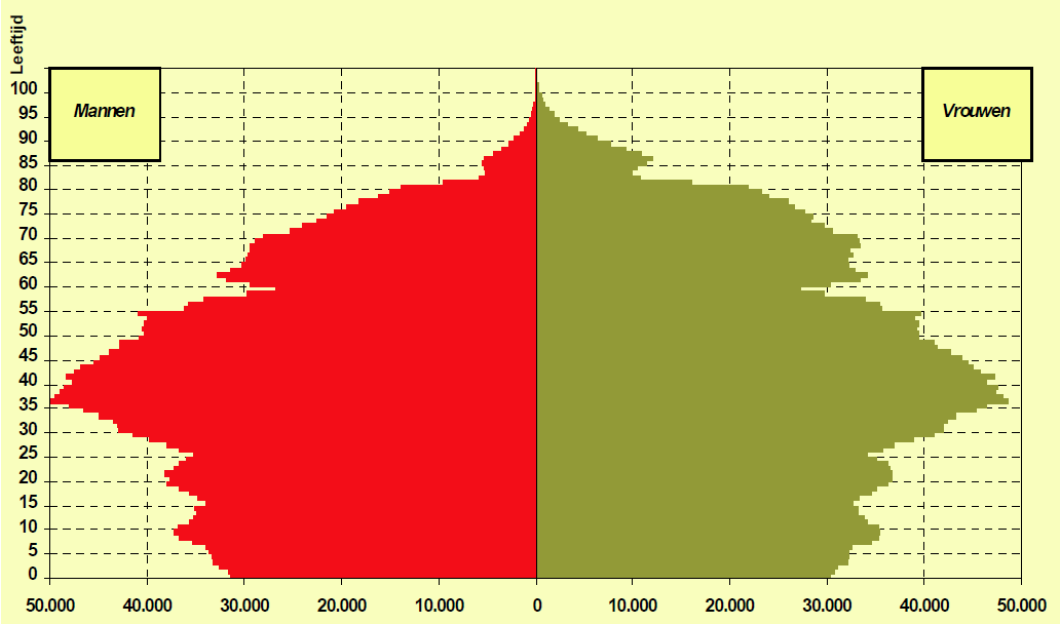
Aandachtspunten

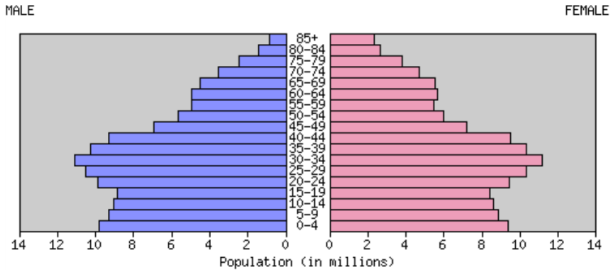
- Omdat je de veranderlijke als continu behandelt, teken je op de x-as de intervallen tegen elkaar. Dit is dus anders dan bij een staafdiagram.
- Een histogram is een "volle" figuur van aaneensluitende rechthoeken waar de oppervlakte de aandacht trekt. Een histogram gebruik je om een globaal zicht te krijgen op continue data. Je kijkt naar kenmerken zoals symmetrie, scheefheid, clusters, … Stap af van de drang om naar hoogtes te kijken maar laat je aandacht trekken door de oppervlakte en de globale vorm.
- Niet alle klassen moeten even breed zijn, zolang je de basiseigenschap (oppervlakte!) van het histogram respecteert.

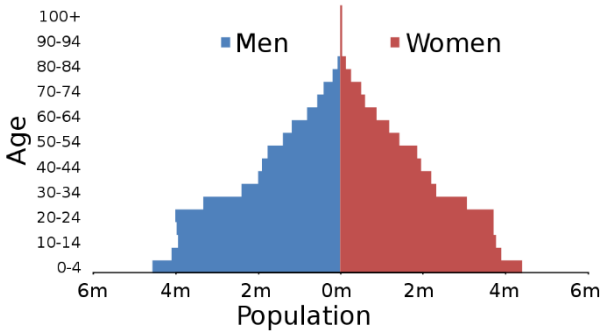
Speciaal type

Een populatiepiramide wordt vaak gebruikt om de bevolking van een land weer te geven. Men maakt dan twee histogrammen, één voor vrouwen, één voor mannen. Elk histogram geeft aan (in aantal of in percent) hoeveel mensen in een bepaalde leeftijdscategorie (per jaar of per interval van 5 jaar) terechtkomen. Die histogrammen worden dan vertikaal tegen elkaar geplaatst.

Hieronder zie je een populatiepiramide voor Vlaanderen, voor de VS (links) en voor Egypte (rechts).





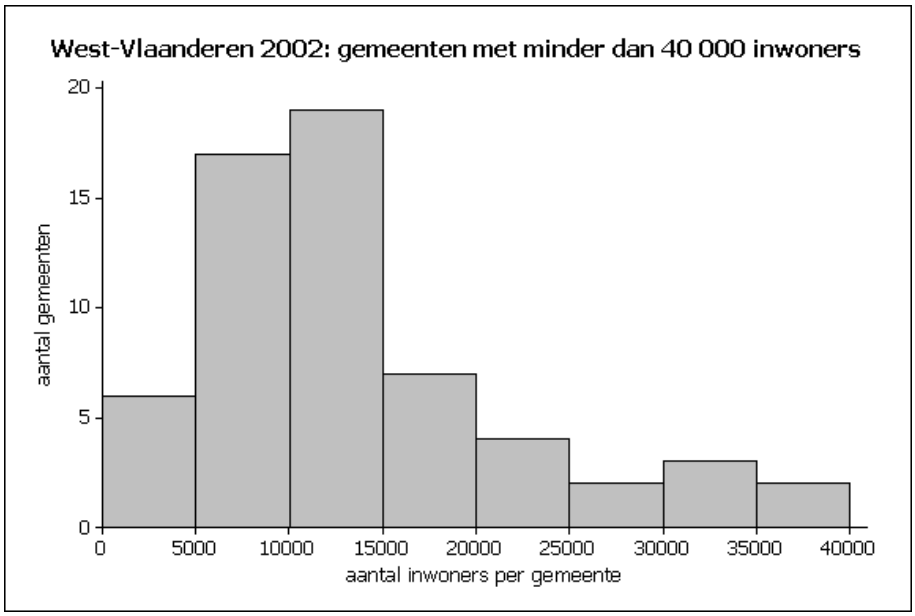


4.1.2. Histogram voor een discreet numerieke veranderlijke

Als je een overzicht wil hebben van het aantal inwoners per gemeente in West-Vlaanderen dan kan je dat mooi illustreren met een dotplot. Dat heb je hierboven gezien.

Qua inwonersaantal springen Brugge, Kortrijk, Oostende en Roeselare duidelijk in het oog. Maar die 60 andere West-Vlaamse gemeenten, hoe beschrijf je die? Je zou je hierbij de vraag kunnen stellen: "Is de verdeling van het aantal inwoners bij die gemeenten ongeveer symmetrisch, met weinig zeer kleine gemeenten, een meerderheid middelgrote en weinig grote?".

Bij deze studie is de naam van de gemeente niet belangrijk, je kijkt alleen naar het aantal inwoners. Dat aantal is een discreet numerieke veranderlijke met heel veel mogelijke waarden. Daarom kan je hier methoden van continu numerieke veranderlijken gebruiken. Je verdeelt het interval [0 ; 40 000[in klassen en telt hoeveel gemeenten in die klassen terechtkomen. Het resultaat kan je dan grafisch voorstellen in een histogram zoals hieronder.



Het histogram is scheef naar rechts. Voor de gemeenten in deze studie is het gemiddeld aantal inwoners 13 644 terwijl de mediaan 11 282 is. Dat is logisch bij een grafiek die "scheef naar rechts" is. Daarom kan je zeggen dat, voor deze groep van 60 gemeenten, een "typische" gemeente 11 282 inwoners telt. Er zijn 30 gemeenten met minder dan 11 282 inwoners en 30 die er meer hebben.

4.1.3. Opdrachten histogram

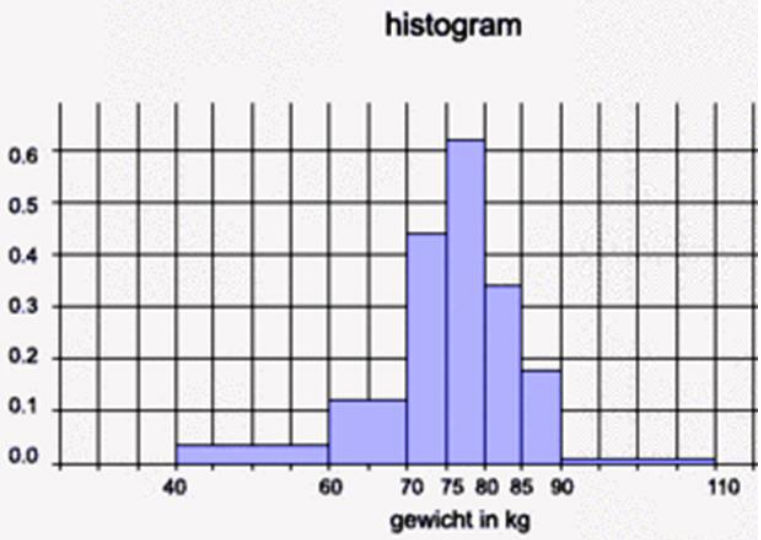
VOORBEELD UIT DE MEDIA

Het gewicht van volwassen vrouwen
http://nl.wikipedia.org/wiki/Histogram 09/03/2012

Korte beschrijving van het onderzoek

Van 100 volwassen vrouwen is het gewicht (in kg) genoteerd. De opmetingen zijn daarna gegroepeerd in klassen. Dit levert de frequentietabel met klassenindeling:

klasse	frequentie
[40 ; 60[	7
[60 ; 70[	12
[70 ; 75[	22
[75 ; 80[	31
[80 ; 85[	17
[85 ; 90[	9
[90 ; 110[	2



Opdracht 11

1. De gewichten zijn afgerond tot op een kilogram. Eigenlijk bestaan de uitkomsten dus uit 100 gehele getallen. Toch wordt die dataset hier grafisch voorgesteld met een histogram. Wat is hiervoor de verklaring?

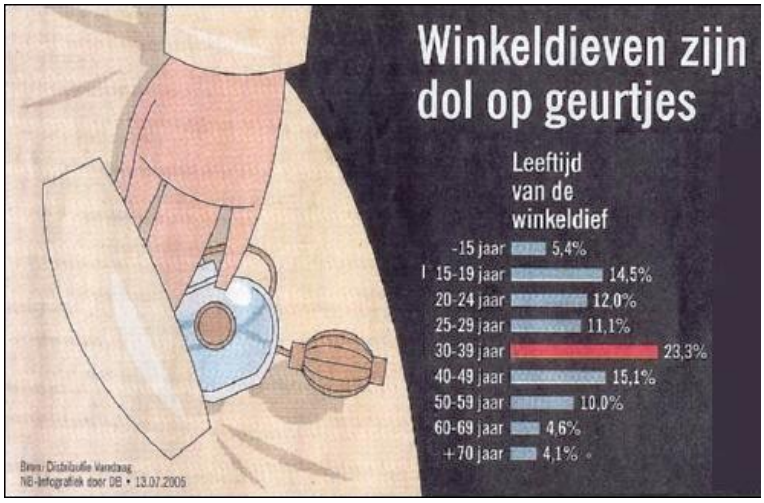
2. De hoogte van de getekende rechthoeken (kijk naar de schaal op de y-as) is, in volgorde: 0.035 ; 0.12 ; 0.44 ; 0.62 ; 0.34 ; 0.18 ; 0.01. Bereken nu de oppervlakte van deze rechthoeken. Zijn de hoogten recht evenredig met de frequentie of zijn het de oppervlakten? Welke basiseigenschap van histogrammen is hier geïllustreerd?

VOORBEELD UIT DE MEDIA

Winkeldiefstallen
Het Nieuwsblad 14/07/2005

Korte beschrijving

Het aantal winkeldiefstallen in ons land is de jongste jaren merkbaar gedaald. Dat blijkt uit gegevens die het Belgisch Comité voor de Distributie in 658 verkooppunten verzamelt. Uit dat onderzoek blijkt ook dat reukwaren het populairst zijn bij winkeldieven.



Hoe zou het onderzoek gevoerd kunnen zijn?

De elementen in de dataset zijn de winkeldieven die (op heterdaad) betrapt zijn en die bijgevolg ondervraagd konden worden. De leeftijd van de dief (in jaren) is hier de bestudeerde veranderlijke.

Opdracht 12

1. Soms blader je eens vluchtig door een krant, zonder echt te lezen. Als je op die manier deze grafiek ontmoet, wat is dan je eerste snelle indruk (in welke leeftijdscategorie zitten de meeste winkeldieven)?

2. Je ziet een figuur waar alle staafjes even breed zijn. Zij staan ook allemaal even ver van elkaar. Dus stellen zij allemaal een even grote leeftijdscategorie voor. Of niet soms?

3. Het rode staafje staat op een leeftijdsgroep van dertigers (winkeldieven die minstens 30 jaar zijn maar jonger dan 40). Dat is een interval van 10 jaar. Kan je uit de gegevens van de grafiek een ander leeftijdsinterval van 10 jaar halen waar er nog meer winkeldieven zijn? Waarom is daar dan niet de aandacht op getrokken met een rode kleur?

4. De bestudeerde veranderlijke is “leeftijd”. Dat is een continue veranderlijke. Welk type grafiek gebruik je om zo’n veranderlijke voor te stellen?

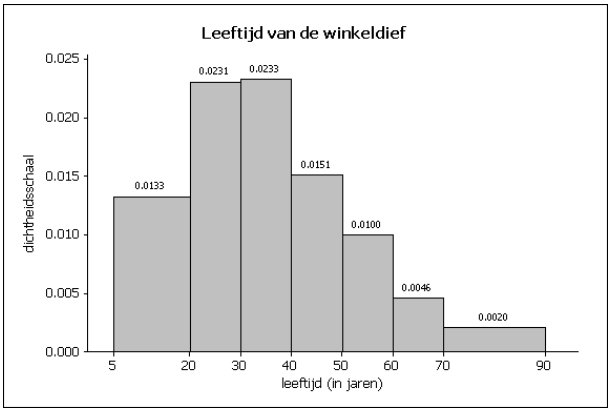
5. Maak nu een juiste grafiek op basis van de gegevens die je van de figuur kan aflezen. Veronderstel daarbij dat winkeldieven niet jonger dan 5 en niet ouder dan 89 zijn. De som van alle percenten is 100.1 %. Je kan er dus van uitgaan dat er geen echte meetfouten zijn gemaakt maar dat 100.1 % ontstaan is door afrondingen bij berekeningen. Daarom mag je de percenten gebruiken zoals ze op de figuur staan.

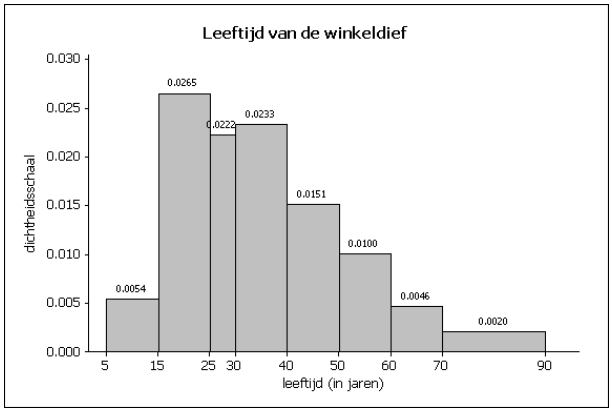
Nota.

Er is geen unieke manier om histogrammen te tekenen. De keuze van de intervallen heeft een invloed op de interpretatie van de gegevens.

Links bestrijken alle intervallen een leeftijd van 10 jaar, behalve de eerste en de laatste. Niet de hoogte maar de oppervlakte van de eerste rechthoek valt hier op. Van alle winkeldieven is $(15)\times(0.0133) \cong 20$ % jonger dan 20 jaar.

Rechts bestrijken, op twee na, alle intervallen een leeftijd van 10 jaar. Van alle winkeldieven is $(10)\times(0.0265) = 26.5$ % tussen 15 en 25 jaar en $(10) \times (0.0233) = 23.3$ % tussen 30 en 40.





5. Twee veranderlijken – beide categorisch

Nota over het aantal veranderlijken

Sommige studies kan je zowel bij één veranderlijke als bij twee veranderlijken onderbrengen.

Om de dotplot van het aantal inwoners van West-Vlaamse gemeenten te tekenen kan je op twee manieren te werk gaan:

1. Je maakt een dataset waarbij de elementen alle inwoners van West-Vlaanderen zijn. Bij elke inwoner noteer je één ding: de woonplaats. Zo werk je met een nominale veranderlijke die 64 verschillende waarden heeft. Voor elke waarde tel je hoeveel keer zij voorkomt. Daarmee heb je genoeg informatie om die dotplot te tekenen.
2. Je maakt een dataset waarbij de elementen de 64 gemeenten van West-Vlaanderen zijn. Bij elke gemeente noteer je de waarde van twee veranderlijken: de naam van de gemeente en het aantal inwoners. Die twee veranderlijken heb je nodig om de dotplot te tekenen.

In sommige teksten lees je over "één veranderlijke met subtypes". De eigenlijke studie gaat bijvoorbeeld over de verdeling van de bloedgroepen. Je voert die studie uit in Vlaanderen maar ook in Japan. Die twee studies vergelijk je dan. In feite werk je hier met een dataset waarbij de elementen mensen zijn waarbij twee veranderlijken werden opgemeten: hun bloedgroep (nominaal) en of ze Vlaming of Japanner zijn (nominaal).

5.1. Staafdiagram met subtypes

5.1.1. Staafdiagram met subtypes bij een nominale veranderlijke

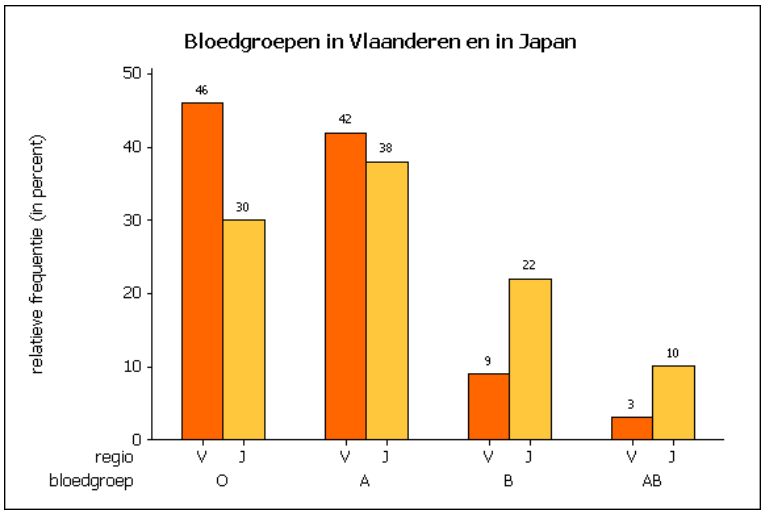
Voorbeeld

In Vlaanderen is bij 200 mensen de bloedgroep bepaald. Het resultaat is:

O	A	B	AB	Totaal
92	84	18	6	200
46%	42%	9%	3%	100%

Eenzelfde studie is uitgevoerd bij 50 Japanners, met resultaat:

O	A	B	AB	Totaal
15	19	11	5	50
30%	38%	22%	10%	100%



Het aantal deelnemers is in Vlaanderen groter dan in Japan. Door de resultaten uit te drukken in percent kan je ze toch grafisch vergelijken in eenzelfde figuur. Dat zie je hier bij het staafdiagram met subtypes. De bestudeerde veranderlijke is de bloedgroep (categorisch). De subtypes worden bepaald door een andere categorische veranderlijke: de regio (Vlaanderen of Japan).

Kenmerken van de grafiek

- Een staafdiagram met subtypes volgt grotendeels de regels van een gewoon staafdiagram.
- Per waarde van de bestudeerde veranderlijke (per bloedgroep) worden nu meerdere staafjes weergegeven (een staafje voor elk subtype). Die staafjes mag je tegen elkaar tekenen.
- Bij het kiezen van de volgorde waarin je de waarden (de bloedgroepen) op de x-as plaatst, kijk je naar één van de subtypes. In dit voorbeeld zijn de bloedgroepen geordend in dalende relatieve frequentie voor Vlaanderen.

5.1.2. Staafdiagram met subtypes bij een ordinale veranderlijke

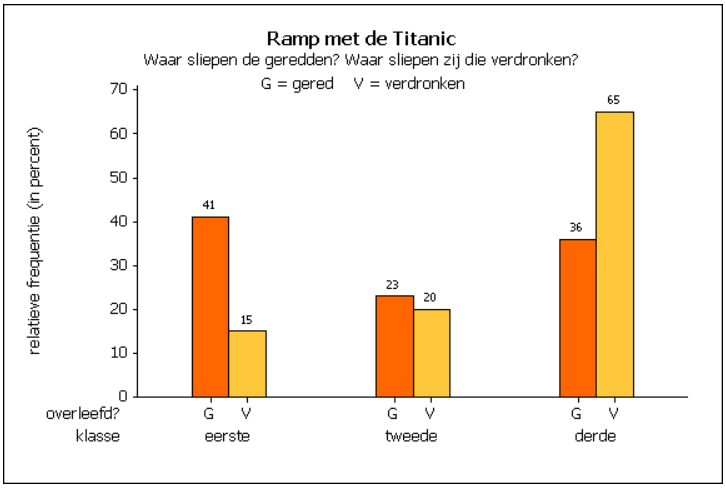
Voorbeeld

Op de Titanic sliepen de passagiers in eerste, tweede of derde klasse. Bij de ramp verdronken meer dan 800 opvarenden.

In de tabel zie je hoe de slachtoffers verdeeld waren over de verschillende passagiersklassen. Je kan die verdeling vergelijken met de groep die gered werd: in welke passagiersklasse sliepen zij?

		Overleefd?	
		gered	verdronken
Passagiers-klasse	eerste klasse	41 %	15 %
	tweede klasse	23 %	20 %
	derde klasse	36 %	65 %
	Totaal	100 %	100 %

Bij de vraag die je hier stelt, bestudeer je de categorische veranderlijke *passagiersklasse*. Die veranderlijke heeft de waarden: eerste klasse (de grootste luxe en hoogste prijs), tweede klasse (wat mindere luxe en een minder hoge prijs) en derde klasse (de minste luxe en het goedkoopste tarief). Er is een logische volgorde (qua luxe en prijs) bij die passagiersklassen. Die volgorde is in de grafiek gerespecteerd.



De passagiersklasse bestudeer je hier bij twee groepen (*gered – verdronken*). Het resultaat van de volledige studie kan je grafisch voorstellen door een staafdiagram met subtypes.

Kenmerken van de grafiek

- De volgorde waarin je de waarden (eerste, tweede, derde klasse) op de x-as plaatst, wordt bepaald door de logische volgorde van de passagiersklassen.
- Verder lijkt dit staafdiagram goed op een staafdiagram met subtypes voor een nominale veranderlijke.

5.1.3. Opdracht staafdiagram met subtypes

VOORBEELD UIT DE MEDIA

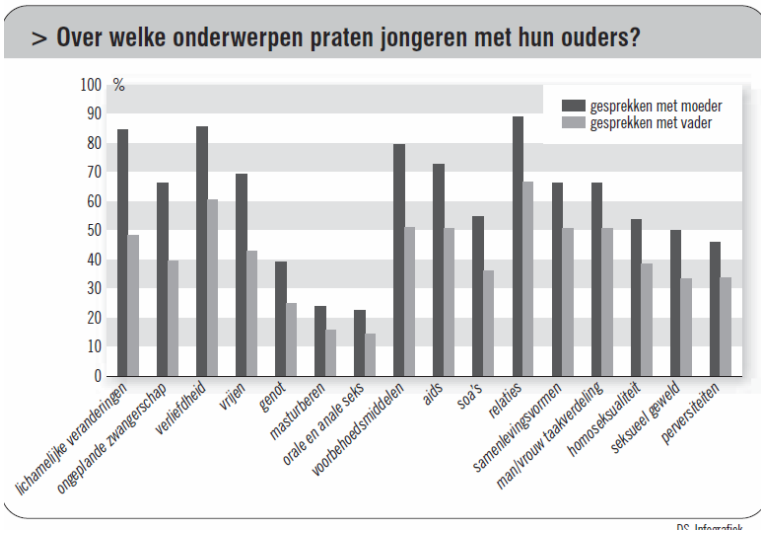
Over welke onderwerpen praten jongeren met hun ouders?
De Standaard 9/03/2005

Korte beschrijving

Hoe hechter de band in het gezin, hoe vaker ouders met hun kinderen over seksualiteit en relaties praten.

Ouders van deze generatie vinden sowieso dat zij dat redelijk vaak doen. Vergeleken met de stilte waarin ze zelf zijn grootgebracht, is dat ook zo.

Jongeren zien het anders. Zij halen de meeste informatie uit andere kanalen, zoals het internet, tv-programma's en tijdschriften.



Hoe zou het onderzoek gevoerd kunnen zijn?

Men heeft aan jongeren een vragenlijst gegeven waarin gevraagd werd of ze al dan niet praten met vader en/of moeder over lichamelijke veranderingen, verliefdheid,… Die lijst bestond uit 16 topics en bij elke topic was er een vakje voorzien voor "vader" en een vakje voor "moeder". In totaal waren er dus 32 vakjes die de ondervraagde kon blanco laten (als er niet werd gepraat) of aankruisen (als er wel werd gepraat). Van de ondervraagde jongeren heeft 85 % aangekruist dat ze met hun moeder praten over lichamelijke veranderingen.

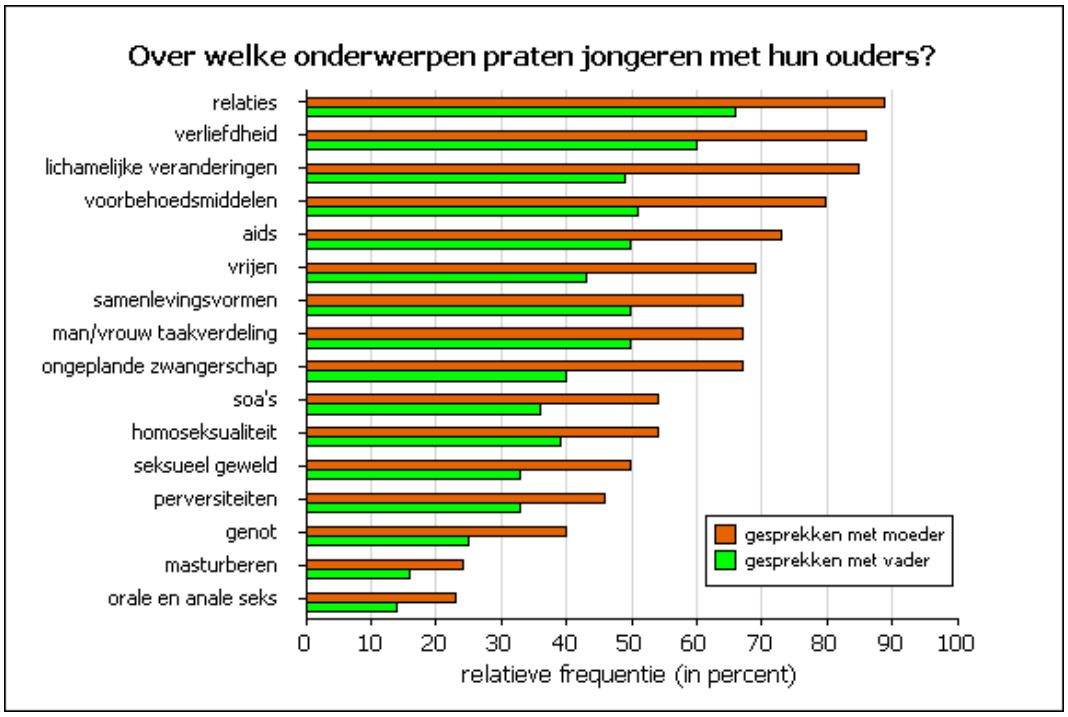
Opdracht 13

De grafiek ziet eruit als een staafdiagram met subtypes. Er is een staafdiagram voor "moeder" en een staafdiagram voor "vader". Die twee staafdiagrammen zijn samengebracht in één grafiek waar per categorie de staafjes van "moeder" en "vader" tegen elkaar zijn getekend (in een verschillende kleur).

1. Het staafdiagram voor "moeder" is geen staafdiagram in de klassieke zin. Hoe zie je dat? Wat is hier gebeurd?

2. De grafiek geeft informatie over *aangekruiste antwoorden*. Je ziet dat er bij elk onderwerp meer antwoorden zijn aangekruist bij "moeder" dan bij "vader". En bij "moeder" zijn er dubbel zoveel kruisjes gezet bij "voorbehoedsmiddelen" dan bij "genot". Dat zie je duidelijk op de figuur want de oorsprong van de y-as is zichtbaar zodat de staafjes een juist beeld geven. Informatie over *de deelnemers* is moeilijker te achterhalen. Veronderstel eens dat je een grafiek zou hebben waarbij je ziet dat 50 % van de jongeren met hun moeder praat over homoseksualiteit en 50 % van de jongeren met hun vader. Kunnen dan de volgende beweringen juist zijn? Geef een voorbeeld om volgende mogelijke antwoorden te staven:
    a. alle ondervraagde jongeren praten thuis over homoseksualiteit
    b. de helft van de ondervraagde jongeren praat thuis nooit over homoseksualiteit.

3. Over seksualiteit en relaties praten jongeren meer met hun moeder dan met hun vader. Maar als je op zoek gaat naar de top 5 van onderwerpen waarover jongeren met hun moeder praten, dan is dat uit de grafiek niet zo eenvoudig op te maken. Daarom kan je de grafiek verbeteren door te sorteren volgens dalend percent bij de staafjes voor "moeder". In onderstaande figuur is dat zo gebeurd. De grafiek is horizontaal geplaatst om de namen van de categorieën makkelijker voluit te schrijven. Wat is nu de top 5 van onderwerpen waarover men met moeder praat?

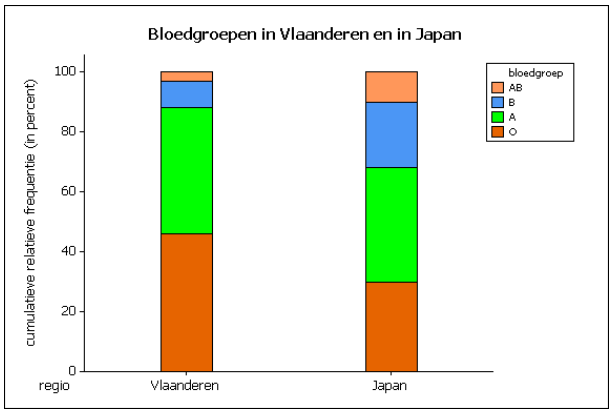


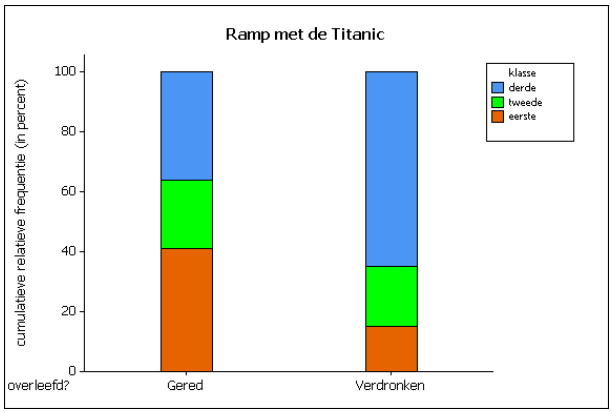
5.2. Gestapeld staafdiagram

In bovenstaande studies vergeleek je twee groepen voor een bepaalde categorische veranderlijke. De ene keer ging het over Vlamingen en Japanners die je vergeleek qua bloedgroep (O, A, B, AB). De andere keer ging het over "wie gered werd" en "wie verdronk" waarbij je keek in welke passagiersklasse zij sliepen. Beide studies werden voorgesteld door een staafdiagram met subtypes.

Je kan ook anders tewerk gaan en per groep één staafje tekenen. Dat staafje stelt dan de hele groep (100 %) voor. Je verdeelt dat staafje in stukken volgens de waarden van de categorische veranderlijke. Als er bij de Vlamingen 46 % bloedgroep O hebben, 42 % bloedgroep A, 9 % bloedgroep B en 3 % bloedgroep AB, dan stapel je 4 staafjes van gepaste grootte op elkaar. In totaal heb je dan alle onderzochte Vlamingen en dat is 100 % wat die groep betreft.

De grafieken voor de bloedgroepen en voor de ramp met de Titanic zien er dan als volgt uit:





Kenmerken van de grafiek

- Bovenstaande grafieken zijn gestapelde staafdiagrammen. Per subtype is er geen gewoon staafdiagram (met staafjes los van elkaar) getekend maar al die staafjes zijn op elkaar geplaatst. Elk subtype (elke groep) krijgt op de x-as één staaf van 100 % die je in de y-richting in stukken verdeelt. De grootte van die stukken toont de proportie waarmee de waarden van de bestudeerde categorische veranderlijke in die groep voorkomen.
- Voor elke staaf neem je, bij het verdelen in stukken, dezelfde volgorde van waarden van de categorische veranderlijke.
- Je gebruikt verschillende kleuren om de verschillende waarden van de categorische veranderlijke aan te geven.

Aandachtspunt

- Hoewel gestapelde staafdiagrammen vaak gebruikt worden, is het dikwijls moeilijk om de grafiek correct te lezen. Op de grafiek waar de staafjes naast elkaar getekend zijn, zie je duidelijk dat bloedgroep A minder voorkomt in Japan dan in Vlaanderen. Bij het gestapelde staafdiagram moet je de groene rechthoeken met elkaar vergelijken. Dat is veel moeilijker.

5.3. Grafieken bij een kruistabel

Soms kan je een studie van twee categorische veranderlijken voorstellen in een kruistabel.

Voorbeeld

Toen de onzinkbaar geachte Titanic op 14 april 1912 tegen een ijsberg opbotste, kwamen veel mensen om het leven. Achteraf bleek dat vooral de armste passagiers tot de slachtoffers behoorden. In onderstaande kruistabel staat aangegeven in welke klasse de passagiers van de Titanic een kajuit hadden en of zij de ramp overleefden.

		Overleefd?		
		gered	verdronken	Totaal
Passagiers-klasse	eerste klasse	203	122	325
	tweede klasse	118	167	285
	derde klasse	178	528	706
	Totaal	499	817	1316

Kenmerken van de tabel

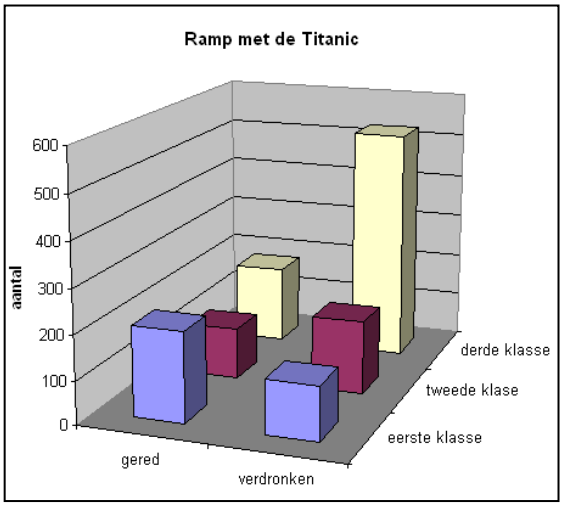
- Horizontaal plaats je de waarden van een eerste categorische veranderlijke.
- Verticaal plaats je de waarden van de andere categorische veranderlijke.
- In de tabel die zo ontstaat, zet je op de snijding van de i-de rij en de j-de kolom het aantal elementen dat tegelijk behoort tot categorie i van de eerste veranderlijke én tot categorie j van de tweede veranderlijke.

Aandachtspunten

- Een kruistabel geeft informatie over de samenhang tussen 2 categorische veranderlijken.
- Een kruistabel geeft ook informatie over elke veranderlijke afzonderlijk. Dat zie je als je aan de tabel rij- en kolomtotalen toevoegt. Wat bijvoorbeeld de passagiersklasse betreft, reisden er $325/1316 \cong 25\,\%$ passagiers in eerste klasse, $285/1316 \cong 22\,\%$ in tweede klasse en $706/1316 \cong 53\,\%$ in derde klasse.

5.3.1. Volledige informatie: een staafdiagram in 3D

Sommige computerprogramma's kunnen een kruistabel weergeven in drie dimensies. De categorieën worden aangegeven in het (x,y)-vlak, het aantal elementen wordt weergegeven door een balkje loodrecht op het (x,y)-vlak.



Kenmerken van de grafiek

- De grafiek heeft een x-, y- en z-as.
- Op de x- en de y-as plaats je de waarden van de categorische veranderlijken.
- Op de z-as komt de absolute (of relatieve) frequentie. De hoogte van elk balkje wordt bepaald door de bijhorende frequentie.

Aandachtspunten

- Een driedimensionale figuur is niet altijd duidelijk. De hoogte van de balkjes is moeilijk afleesbaar en soms verbergen balkjes andere balkjes.
- Bepaalde informatie zoals "*het percent overlevenden per passagiersklasse"* lees je niet zomaar af uit een 3D figuur. Wat je dan moet doen, ontdek je in volgend punt.

5.3.2. Voorwaardelijke informatie per rij of per kolom

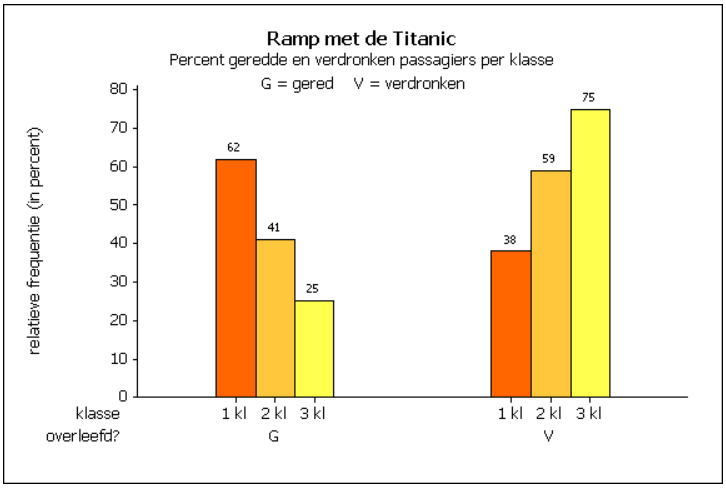
Bij een kruistabel kan je voorwaardelijk werken. Je conditioneert dan op een waarde van één van de categorische veranderlijken. Bij *passagiersklasse* kan je conditioneren op *eerste klasse.* Dat betekent dat je alleen kijkt naar de passagiers die in eerste klasse reisden. Dat is nu je nieuwe groep. Van die 325 mensen werden er 203 gered. De proportie $203/325 \cong 0.62$ (of 62 percent) is een voorwaardelijke proportie (of een voorwaardelijk percent), voorwaardelijk op de passagiers in *eerste klasse*. Je kan zo bij elke klasse tewerk gaan en dan krijg je de tabel hiernaast.

		Overleefd?		
		gered	verdronken	Totaal
Passagiers-klasse	eerste klasse	203 = 62 %	122 = 38 %	325 = 100 %
	tweede klasse	118 = 41 %	167 = 59 %	285 = 100 %
	derde klasse	178 = 25 %	528 = 75 %	706 = 100 %

In de eerste klasse is 62 % van de passagiers gered en is 38 % verdronken. Je kan dit voorstellen met een staafdiagram waar je twee staafjes tekent die los van elkaar staan.

Eenzelfde studie kan je doen voor de tweede en voor de derde klasse. Samen zijn dat 3 studies die elk afzonderlijk sommeren tot 100 %.

Om je resultaten grafisch voor te stellen teken je een staafdiagram met subtypes.



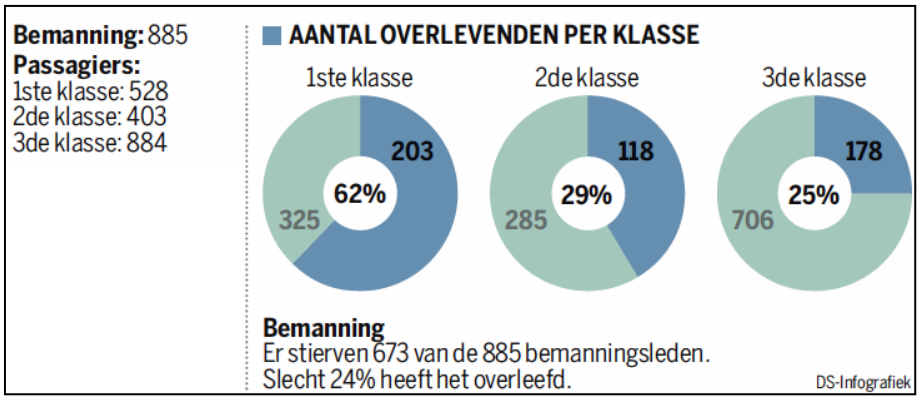
5.3.3. Opdracht studie met subtypes: de Titanic

Voor de passagiers op de Titanic kan je, per passagiersklasse, een studie maken over het percent geredden in die klasse. Hierboven is dat gebeurd in een tabel waar je de aantallen ziet samen met de voorwaardelijke proporties per passagiersklasse. Die proporties zijn ook grafisch voorgesteld in een staafdiagram met subtypes.

VOORBEELD UIT DE MEDIA

Er was niets mis met de Titanic
De Standaard 23/03/2012

In de media ontmoet je een variëteit van grafische voorstellingen om eenzelfde gebeurtenis voor te stellen. Hieronder zie je zo'n voorbeeld.



Opdracht 14

De krant vermeldt ook de bemanningsleden maar de studie gaat hoofdzakelijk over de passagiers. In de linkerkolom staat het totaal aantal passagiers per klasse. Rechts staan 3 taartdiagrammen. Zij starten allemaal op "twaalf uur" en houden zich aan eenzelfde volgorde: eerst zij die overleefden (blauw) en dan zij die verdronken (groen). Op die manier kan je de overlevenden per klasse goed vergelijken. Bij een staafdiagram met subtypes staan de staafjes van de geredden naast elkaar. Dat is makkelijker te vergelijken dan hier, waar je ogen heen en weer over de drie taartdiagrammen moeten springen om de grootte van de blauwe sectoren in te schatten.

1. Bestudeer het taartdiagram van de eerste klasse. Wat betekenen de drie getallen? Komen die overeen met de getekende sectoren? Wat loopt er mis?

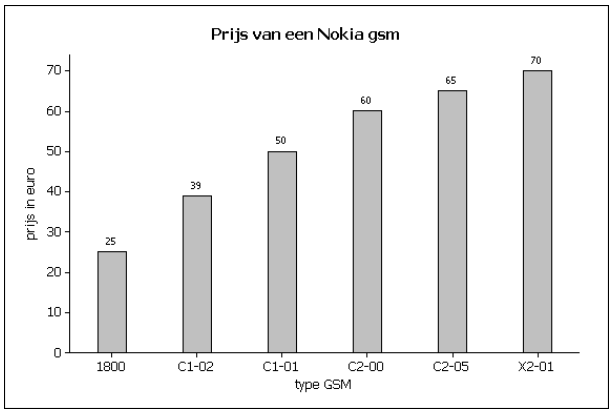
2. Bestudeer het taartdiagram van de tweede klasse. Wat betekenen de drie getallen? Komen die overeen met de getekende sectoren? Wat loopt er mis?

3. Als je het taartdiagram van de tweede en de derde klasse met elkaar vergelijkt, dan zie je een redelijk grote blauwe sector waarin het getal 118 staat en een veel kleinere blauwe sector waarin het getal 178 staat. Dat lijkt een perfecte manier om de lezer op het verkeerde been te zetten. Wat kan hier beter als je dan toch met taartdiagrammen wil werken? Teken een betere figuur.

5.4. Staafdiagram voor twee veranderlijken

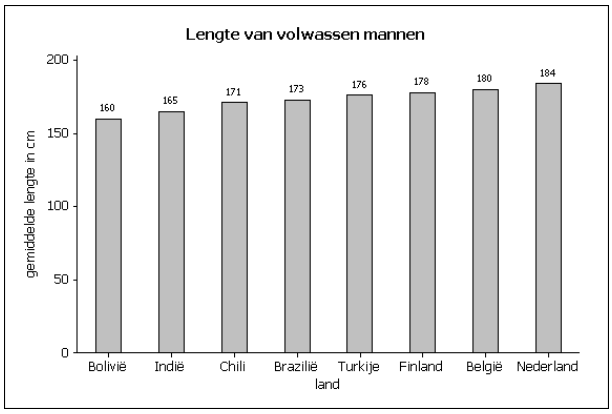
Voorbeeld

Bij een "klassiek" staafdiagram toon je de frequentie (of relatieve frequentie) van elke waarde van één categorische veranderlijke. Maar je kan ook, bij elke waarde van een categorische veranderlijke, de waarde van een andere (numerieke) veranderlijke tonen.



Die numerieke veranderlijke kan discreet zijn, zoals de prijs van een gsm. Hiernaast zie je een voorbeeld van prijzen voor verschillende types van goedkope Nokia gsm's.

De numerieke veranderlijke kan ook continu zijn, zoals de lengte van volwassen mannen. Van een continu numerieke veranderlijke toon je, per categorie, één getal (gewoonlijk een kengetal zoals het gemiddelde). Hiernaast zie je voor een aantal landen de gemiddelde lengte van hun mannelijke inwoners.

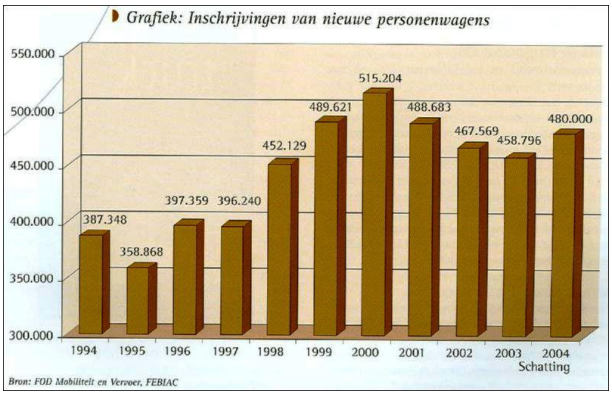


VOORBEELD UIT DE MEDIA

Inschrijvingen van nieuwe personenwagens
Febiac 1/01/2005

Korte beschrijving

Het economisch herstel dat midden 2003 werd ingezet, werd in de loop van 2004 bevestigd. De groei werd in hoofdzaak gestuurd door een toenemende binnenlandse vraag (+3.2%) en een hogere privéconsumptie (+2.2%). Gelijklopend hiermee zijn ook de investeringen door bedrijven significant toegenomen (+3.5%). Het vertrouwen van de consument en van de ondernemingen heeft zich sterk hersteld in vergelijking met 2003. Ten gevolge daarvan was 2004 een uitstekend autojaar.



Hoe zou het onderzoek gevoerd kunnen zijn?

Als je je auto inschrijft, dan wordt daarvan een dataset met type auto, nummerplaat, jaartal, datum,… bijgehouden. Je kan nu een deelstudie maken en je beperken tot 1 veranderlijke: het jaartal. Je telt dan hoeveel keer een bepaald jaartal voorkomt bij alle nieuwe personenwagens die ingeschreven zijn van 1994 tot 2004. Zo ken je, per jaar, het aantal nieuwe wagens. Je kan hierbij het jaartal als een discreet numerieke veranderlijke behandelen.
Voor 2004 staat er "schatting". Waarschijnlijk was de volledige dataset voor 2004 nog niet beschikbaar (misschien is de grafiek in november 2004 gemaakt), maar had men toch al een idee over het uiteindelijke aantal.

Opdracht 15

Om de grafiek te verbeteren, helpt het om eerst de volgende vragen te beantwoorden.

1. Het staafje boven 1999 is dubbel zo groot als dat boven 1997. Dus zijn er in 1999 dubbel zoveel nieuwe auto's ingeschreven als in 1997. Of niet soms?

2. Is het nodig om de staafjes in 3D te tekenen? Helpt dat om de grafiek te verduidelijken of is dat eigenlijk overbodig (en misschien zelfs verwarrend)?

3. Teken nu een betere grafiek.

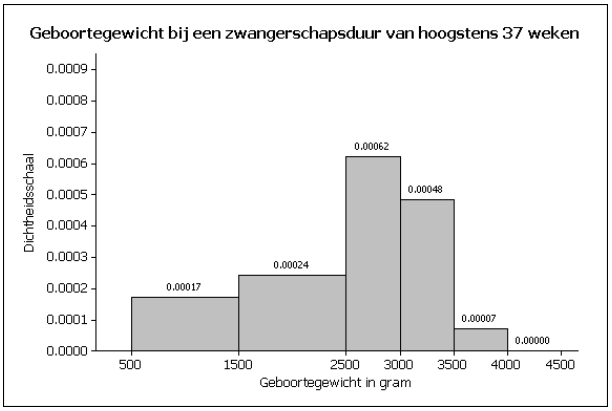
6. Twee veranderlijken – één categorisch en één continu

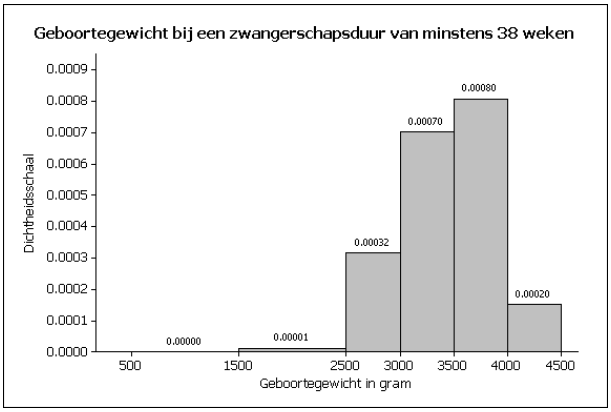
6.1. Histogram op de dichtheidsschaal

De basiseigenschap van een histogram zegt dat **de oppervlakte** van een rechthoek recht evenredig moet zijn met het aantal observaties in de klasse. Een histogram op de dichtheidsschaal is een histogram waarvan de totale oppervlakte gelijk is aan 1 (of aan 100 %). De oppervlakte van een rechthoek is dan gelijk aan het percent observaties in de klasse waarop die rechthoek staat.

Voorbeeld

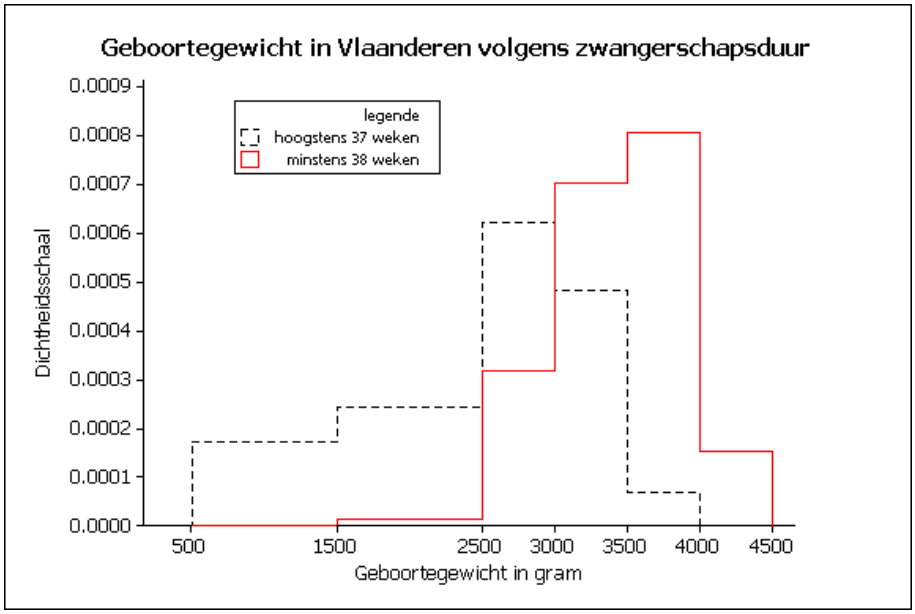
Een steekproef van 200 baby's die in Vlaanderen in het jaar 2000 geboren zijn, leverde 29 baby's waar de zwangerschap hoogstens 37 weken had geduurd en 171 baby's waar de zwangerschap minstens 38 weken had geduurd. Van beide groepen is het geboortegewicht bestudeerd. Hierbij is extra aandacht besteed aan risicobaby's (geboortegewicht kleiner dan 1500 g) en baby's met een laag geboortegewicht (tussen 1500 g en 2500 g). Hieronder zie je een histogram voor beide groepen.





Bij de baby's met een korte zwangerschapsduur behoorde $(1500 - 500) \times (0.00017) = 17$ % tot de groep van risicobaby's. Bij een zwangerschapsduur van minstens 38 weken waren er in deze steekproef geen risicobaby's.

Om de twee groepen goed te kunnen vergelijken teken je de histogrammen op eenzelfde figuur. Je krijgt een duidelijk beeld als je alleen de "omtrek" van de histogrammen tekent.



Kenmerken van de grafiek

- Het "geboortegewicht" is een continue veranderlijke die kan voorgesteld worden door een histogram.
- De "zwangerschapsduur" is een continue veranderlijke die in dit vraagstuk omgevormd is tot een categorische veranderlijke met 2 categorieën: "hoogstens 37 weken" en "minstens 38 weken".
- Het verband tussen korte of lange zwangerschapsduur en het geboortegewicht is hier voorgesteld door 2 histogrammen op de dichtheidsschaal. Op die manier zijn de twee groepen van gewichten correct vergelijkbaar, ook al zijn zij niet even groot en zijn de klassen niet allemaal even breed.

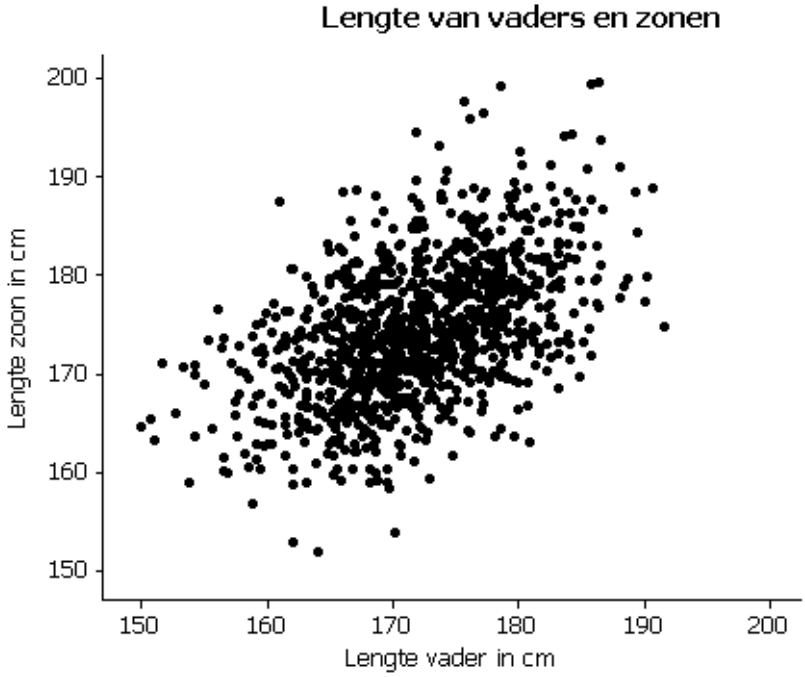
7. Twee veranderlijken – beide continu

7.1. Puntenwolk

Een puntenwolk (of spreidingsdiagram) wordt regelmatig gebruikt in statistiek, maar in de media kom je zo'n figuur zelden tegen.

Voorbeeld

Zoals zijn neef Darwin was Francis Galton (1857–1936) geboeid door genetica. Hij bestudeerde daarbij allerlei kenmerken, ondermeer de lengte van ouders en hun kinderen. Hieronder zie je een puntenwolk die het verband beschrijft tussen de lengte van een vader en van zijn oudste volwassen zoon.



Kenmerken van de grafiek

- Een puntenwolk gebruik je om twee continue veranderlijken in één grafiek weer te geven.
- De gegevenspunten in een puntenwolk worden voorgesteld als koppels (x_i,y_i), waarbij x_i overeenkomt met de eerste veranderlijke (lengte vader) en y_i met de tweede (lengte zoon). De verschillende gegevenspunten worden niet verbonden. Bij één x-waarde kunnen verschillende y-waarden voorkomen.

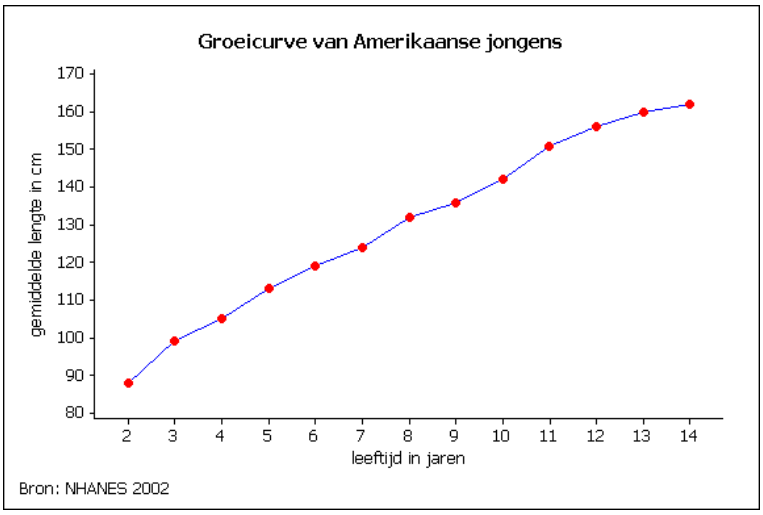
Aandachtspunten

- Bij een puntenwolk is het niet noodzakelijk dat de oorsprong van het assenstelsel zichtbaar is.
- Met een puntenwolk tracht je zicht te krijgen op het verband tussen de twee veranderlijken.

7.2. Lijndiagram

Voorbeeld

De grafiek hiernaast toont een lijndiagram dat de gemiddelde lengte weergeeft van jongens tussen 2 en 14 jaar. De gegevens zijn afkomstig van de Amerikaanse overheid en werden verzameld in het jaar 2002.



Kenmerken van de grafiek

- Een lijndiagram gebruik je om twee continue veranderlijken in één grafiek weer te geven.
- De gegevenspunten in een lijndiagram worden voorgesteld als koppels (x_i,y_i), waarbij x_i overeenkomt met de eerste veranderlijke (leeftijd) en y_i met de tweede (gemiddelde lengte). De verschillende gegevenspunten worden verbonden met een lijnstuk. Bij één x-waarde mag maar één y-waarde voorkomen.

Aandachtspunten

- Bij een lijndiagram is het niet noodzakelijk dat de oorsprong van het assenstelsel zichtbaar is omdat een lijndiagram de aandacht trekt op verandering.
- Vaak geeft een lijndiagram een evolutie in de tijd weer. Men spreekt dan ook van een tijdsgrafiek.

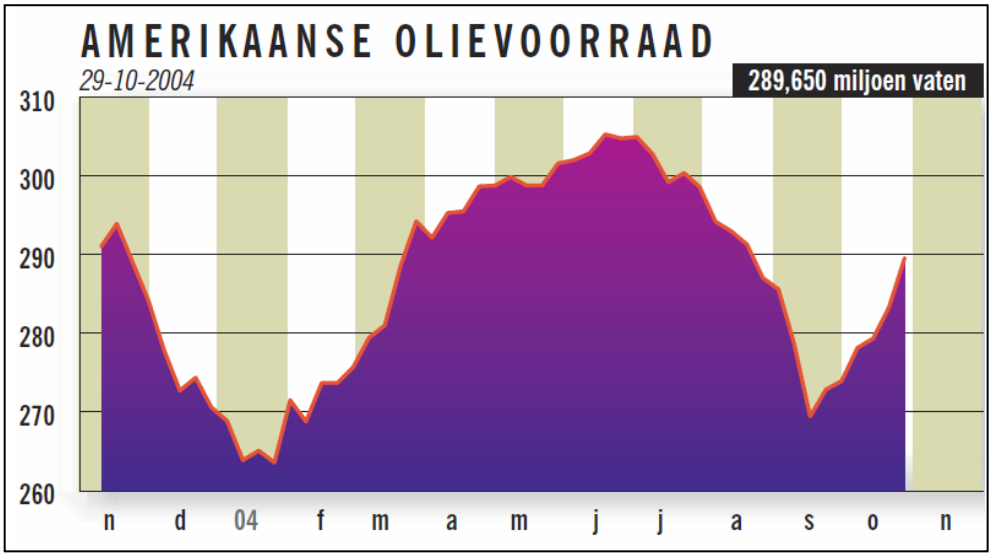
VOORBEELD UIT DE MEDIA

Verdere daling olieprijs heeft meeste aanhangers
De Standaard 8/11/2004

Korte beschrijving

Oliehandelaren en –analisten weten niet goed welke richting de olieprijs de volgende dagen zal uitgaan. Dat blijkt uit een rondvraag van het persagentschap Bloomberg. Een meerderheid verwacht een prijsdaling of een stabilisering van de olieprijs op het huidige peil.

De oliemarkten kijken momenteel vooral naar de stijging van de olievoorraden. Die zijn aan het toenemen.



Hoe zou het onderzoek gevoerd kunnen zijn?

Een mogelijke dataset zou er als volgt kunnen uitzien: de elementen zijn de wekelijkse noteringen van de olievoorraad door de Amerikaanse overheid. De veranderlijken die hierbij genoteerd worden, zijn de datum en de hoeveelheid olievaten, uitgedrukt in miljoenen.

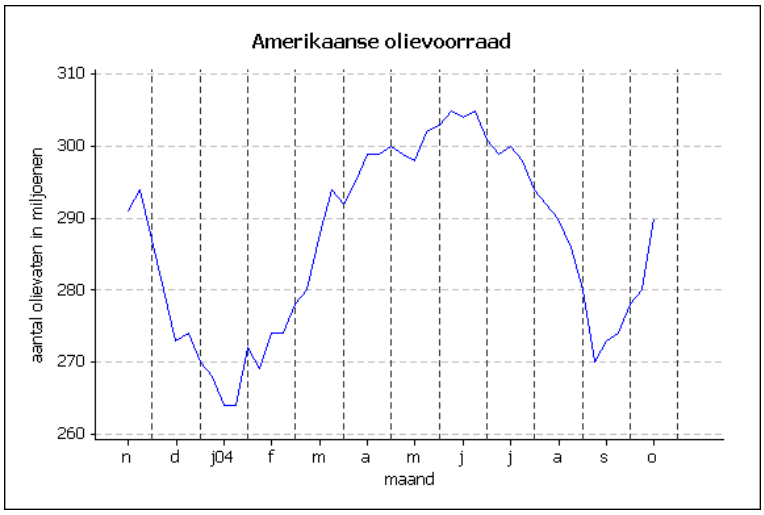
Kenmerken van de grafiek

- De grafiek is een lijndiagram (of tijdsgrafiek), die een evolutie in de tijd weergeeft. De verschillende meetpunten mogen dus met elkaar verbonden worden.
- De grafiek geeft een samenhang weer tussen de datum en de bijhorende olievoorraad voor een periode van november 2003 tot oktober 2004.
- Op de x-as wordt de eerste veranderlijke, de datum (per maand) vermeld. Op de y-as staat de tweede veranderlijke, de olievoorraad (in miljoenen olievaten).
- Op de x-as staan de verschillende periodes correct weergegeven: alle maanden zijn vertegenwoordigd en liggen op gelijke afstand van elkaar.
- De y-as hoeft niet van nul te beginnen als je een verandering (of evolutie in de tijd) wil tonen. Hier is er echter een probleem omdat men de oppervlakte onder de lijn heeft ingekleurd. Zo wordt je aandacht getrokken op een oppervlakte en heb je de neiging om oppervlaktes te vergelijken. Je kan dus beter niet inkleuren en een eenvoudige lijn gebruiken om de aandacht op de veranderende olievoorraad te trekken. Dat zie je hieronder.

Verbeterde grafiek

Om een goed beeld te krijgen van de evolutie die je met een grafiek wil tonen, kan je te werk gaan zoals hiernaast.

De oppervlakte onder de grafiek is nu niet meer ingekleurd waardoor de aandacht meer op de lijn en de evolutie van de olievoorraad wordt getrokken.



VOORBEELD UIT DE MEDIA

Inschrijvingen van nieuwe personenwagens
Febiac 1/01/2005

Opdracht 16

In opdracht 15 heb je een verbeterd staafdiagram getekend voor de grafiek hiernaast. Daarbij was het de bedoeling om op een juiste manier het aantal nieuwe personenwagens per jaar weer te geven.

Veronderstel nu even dat het de bedoeling was om de evolutie van het aantal nieuwe personenwagens over de jaren heen te tonen. Welk grafiektype zou je dan gebruiken? Teken nu zo'n grafiek.