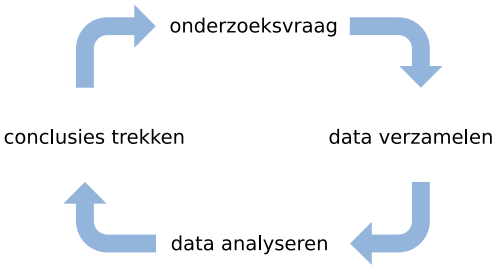
1.1 Statistisch onderzoek doen en data exploreren

Inleiding

Statistiek houdt zich bezig houdt met het verzamelen, uitleggen en presenteren van betrouwbare gegevens over groepen (van bijvoorbeeld mensen). Meestal wordt daarbij maar een deel van de groep onderzocht.

Het woord **statistiek** is afkomstig uit het Latijn: **statisticum collegium**. Dit betekent: les over staatszaken. Je zou kunnen zeggen: de analyse van staatsgegevens.



Figuur 1

Je leert in dit onderwerp

- de begrippen statistisch onderzoek, steekproef, populatie, aselect en representatief;
- onderscheid maken tussen kwalitatieve en kwantitatieve variabelen en tussen discrete en continue kwalitatieve variabelen;
- valkuilen bij het doen van statistisch onderzoek;

Voorkennis

- de basisbegrippen uit de beschrijvende statistiek in de onderbouw vwo;
- de basistechnieken van het werken met de grafische rekenmachine en/of met Excel.

Verkennen

Opgave V1

Files worden meestal ervaren als een probleem.

a Noem minimaal vier bedrijven of beroepsgroepen waarvoor de ontwikkeling van het fileprobleem van belang is.

b Waarom is statistisch onderzoek noodzakelijk om het ontstaan van files te onderzoeken?

c Noem vier (soorten) organisaties die in Nederland statistisch onderzoek doen.

Opgave V2

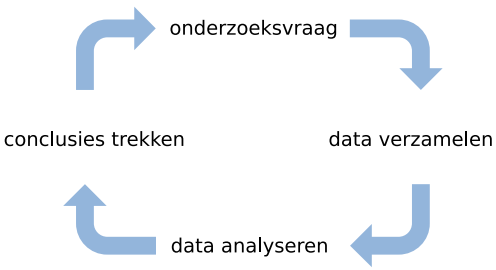
Bedenk een opzet voor een onderzoek naar het gebruik van de fiets onder de ouders en leerlingen van je school. De bedoeling is dat je na het onderzoek uitspraken kunt doen over het gebruik van de fiets van alle leerlingen van de school en hun ouders.

Uitleg 1

Als je kenmerken van een groep mensen of een groep voorwerpen wilt leren kennen, ben je bezig met statistisch onderzoek. Een statistisch onderzoek doorloopt in principe altijd de zogenoemde statistische cyclus.

De eerste stap van de statistische cyclus is het formuleren van een onderzoeksvraag.

Vervolgens moeten er antwoorden (data) op die vraag worden verzameld. Het is vaak onmogelijk of te duur om de volledige groep, de populatie, te onderzoeken. Je neemt dan een steekproef die een weerspiegeling van de populatie moet zijn.



Figuur 2

Een steekproef lijkt in de volgende gevallen op de populatie:

- De steekproef moet representatief zijn. Dit betekent dat alle verschillende elementen van de steekproef in verhouding even vaak in de steekproef moeten voorkomen als in de populatie.
- De steekproef moet ook aselect (willekeurig) zijn. Elk element uit de populatie moet een even grote kans hebben om voor de steekproef gekozen te worden.
- De steekproefomvang moet voldoende groot zijn. Bij een te kleine steekproef zullen de resultaten te veel van toevallige factoren afhankelijk zijn.

Als de gegevens zijn verzameld, moeten ze worden geordend. Dat ordenen gebeurt in tabellen, diagrammen en grafieken. Vervolgens moeten deze gegevens worden geanalyseerd. Daar zijn afhankelijk van het soort gegevens verschillende methoden voor. Uit de analyses van de gegevens kunnen uiteindelijk bepaalde uitspraken over bijvoorbeeld gemiddeldes, waarschijnlijkheid en verwachtingen worden gedaan. Deze uitspraken kunnen eventueel weer worden getoetst en daarbij doorloop je de statistische cyclus nogmaals.

Opgave 1

In welke gevallen is er sprake van een aselecte steekproef? Licht je antwoord toe.

- Je selecteert tien Deventernaren door uit het bevolkingsregister van Deventer de eerste tien namen te nemen die met de letter H beginnen.
- Je kiest een provincie in Nederland door deze geblinddoekt op een kaart van Nederland aan te wijzen.
- Je kiest vijf havo 4-leerlingen door uit een zak met dichtgevouwen lootjes met alle achternamen van leerlingen uit 4 havo zonder kijken de eerste vijf te halen.

Opgave 2

Enkele manieren om data te verwerven zijn weergegeven. In een aantal van die gevallen is de gebruikte methode niet geschikt. Leg telkens uit waarom niet.
Gebruik indien mogelijk de termen aselect en/of representatief.

a Het CBS houdt een enquête over onderwijs op een school voor volwassenenonderwijs.

b Het CBS houdt een enquête over politieke voorkeuren in Nederland. Hiertoe vragen onderzoekers een maand voor de Tweede Kamerverkiezingen in Noord-Brabant aan mensen op straat wat ze gaan stemmen.

c Het CBS houdt in honderd winkelcentra in Nederland een enquête over hoe mensen aankopen doen. Is dit de juiste doelgroep?

Uitleg 2

In een dataset kunnen verschillende soorten gegevens voorkomen. Deze gegevens zijn statistische variabelen. Er wordt een onderscheid gemaakt tussen twee soorten variabelen: kwantitatieve en kwalitatieve variabelen.

Kwantitatieve variabelen zijn variabelen die in getallen zijn uit te drukken. Deze variabelen kun je onderverdelen in discrete variabelen en continue variabelen.

- Discrete variabelen zijn variabelen die geen tussenwaarden kunnen aannemen. Bijvoorbeeld het aantal kinderen in een gezin, een score op een toets van veertig meerkeuzevragen, leeftijd, schoenmaat, enzovoort.
- Continue variabelen zijn variabelen als lengte, gewicht, buitentemperatuur, tijd, enzovoort. Continue variabelen kunnen allerlei tussenwaarden aannemen.

Kwalitatieve variabelen geven een eigenschap of kwaliteit weer, zoals geslacht of politieke voorkeur. Deze variabelen kun je onderverdelen in nominale en ordinale variabelen.

- Nominale variabelen zijn variabelen waarbij het slechts gaat om de naam van datgene wat je wilt meten. Voorbeelden zijn geslacht, politieke voorkeur, provincie, enzovoort.
- Ordinale variabelen kun je ordenen, bijvoorbeeld van laag naar hoog. Voorbeelden van zulke variabelen zijn opleidingsniveau, rang in het leger, eens/neutraal/oneens, enzovoort.

Opgave 3

Geef bij de volgende onderwerpen aan of gaat het om kwalitatieve of kwantitatieve gegevens. Als ze kwalitatief zijn, geef dan ook aan of ze nominaal of ordinaal zijn. Geef bij kwantitatieve gegevens aan of ze discreet of continu zijn:
lengte - plaats van herkomst - hobby - drie favoriete vakantiebestemmingen - aantal zeehonden - een score op een toets van veertig meerkeuzevragen - buitentemperatuur.

Opgave 4

In de Nationale Wetenschapsquiz kwam de vraag voor: “Hoeveel schoolgaande kinderen zijn er gemiddeld per gezin?”
Er wordt een grote steekproef onder schoolkinderen genomen. Er wordt aan hen gevraagd hoeveel schoolgaande broertjes en zusjes ze hebben. Op basis daarvan wordt het gemiddelde aantal schoolgaande kinderen per gezin bepaald.

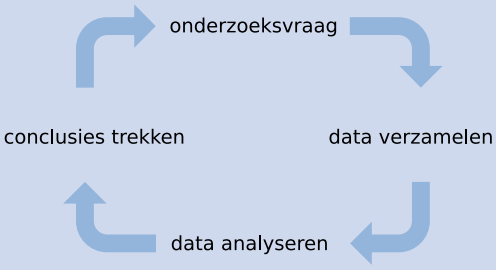
a Heb je hier te maken met een discrete of een continue variabele?

b Kun je verwachten door dit onderzoek een goede inschatting of een te lage of een te hoge schatting van het aantal schoolgaande kinderen per gezin te krijgen?

Theorie en voorbeelden

Om te onthouden

Statistiek houdt zich bezig met het verzamelen, uitleggen en presenteren van betrouwbare gegevens over groepen. In een statistisch onderzoek wordt er gebruikgemaakt van een statistische cyclus.



Figuur 3

De **populatie** is de groep die onderzocht wordt. Dit kunnen mensen zijn, maar ook dieren, voorwerpen of gebeurtenissen. In het algemeen wordt gezegd dat een populatie bestaat uit elementen. Voor het onderzoek wordt een **steekproef** gedaan. Dit is een selectie uit de populatie, omdat het vaak te veel werk is om alle elementen individueel te onderzoeken.

Een steekproef lijkt in de volgende gevallen op de populatie:

- De steekproef moet **aselect** zijn. Dat betekent dat de elementen in de steekproef puur op toevalsbasis uit de populatie gekozen moeten worden.
- De steekproef moet **representatief** zijn. Dat betekent dat bepaalde klassen in de populatie, als daar onderscheid in bestaat, in dezelfde verhouding in de steekproef worden weergegeven.
- Bovendien moet de **steekproefomvang** voldoende groot zijn. Bij een te kleine steekproef zullen de resultaten te veel van toevallige factoren afhankelijk zijn.

Een onderzoek betreft één of meer **statistische variabele(n)**. Leeftijd, geslacht, lengte, gewicht en kleur van ogen zijn variabelen. Er wordt een onderscheid gemaakt tussen twee soorten variabelen: kwantitatieve en kwalitatieve variabelen.

Kwantitatieve variabelen hebben een getalswaarde, er kan mee gerekend worden. Er zijn twee verschillende typen kwantitatieve variabelen:

- **discrete** variabelen, die bestaan uit getallen zonder tussenwaarden;
- **continue** variabelen, die bestaan uit getallen met tussenwaarden.

Kwalitatieve variabelen geven een eigenschap of kwaliteit weer. Er zijn ook twee verschillende soorten kwalitatieve variabelen:

- **nominale** variabelen, dat zijn variabelen die alleen een soort aangeven;
- **ordinale** variabelen, dat zijn variabelen die een soort aangeven, maar die je ook in een bepaalde rangorde kunt plaatsen.

Voorbeeld 1

Bekijk de volgende vier manieren om een steekproef samen te stellen. Geef van elke manier aan of de steekproef representatief, aselect is en of de steekproefomvang groot genoeg is.

1. Voor een onderzoek naar de mening van abonnees op een krant wordt een enquête gehouden. Door middel van loting worden tien abonnees bevraagd.
2. Voor een onderzoek naar het rijgedrag van vrachtwagenchauffeurs enquêteer je mensen. Je kiest voor de uitgang/ingang van een treinstation en bevraagt vanaf 7:00 uur elk uur van de dag tien willekeurige reizigers.
3. Voor een onderzoek naar het rookgedrag van ouders van leerlingen van je school ondervraag je de eerste vijftig binnenkomende ouders op een ouderavond.
4. Voor een onderzoek naar het rookgedrag onder ouders van leerlingen van je school selecteer je door loting vijftig leerlingen van je school en ondervraag je weer na loting de vader of de moeder van elk van de vijftig leerlingen.

Antwoord

Steekproef 1 is aselect, want alle abonnees hebben een even grote kans om in de steekproef te komen. De grootte van de steekproef is te klein en daarom is de steekproef niet representatief.

Bij steekproef 2 is de populatie al niet goed vastgelegd. Het is dus onzinnig om over aselect, representatief en grootte van de steekproef te spreken.

Steekproef 3 is niet aselect, want ouders die niet op de ouderavond zijn, kunnen niet in de steekproef komen. Als er geen verband is tussen ouderavondbezoek en roken, zou deze steekproef wel representatief kunnen zijn. Of de grootte van de steekproef groot genoeg is, is afhankelijk van het aantal leerlingen op de school.

Steekproef 4 is niet aselect. Alleenstaande ouders met één kind op school hebben een grotere kans om in de steekproef te komen dan getrouwde ouders met één kind op school. Of de steekproef representatief is, is twijfelachtig. Of de grootte van de steekproef groot genoeg is, is afhankelijk van het aantal leerlingen op de school.

Opgave 5

Bekijk de beschrijving van enkele statistische onderzoeken. Benoem telkens de populatie, de statistische variabele en omschrijf een aselecte en representatieve steekproef.

a Er wordt vermoed dat er een lineair dalend verband bestaat tussen het gemiddelde aantal uren dat een persoon wekelijks computerspellen speelt, en diens leeftijd (geteld vanaf tienjarigen).

b Er wordt vermoed dat vrouwen gemiddeld meer wijn drinken dan mannen.

Opgave 6

Je doet een onderzoek onder jongeren naar hun mening over smartphones.

Welke van de genoemde onderzoeksmiddelen is het meest geschikt?
Licht je antwoord toe.

A. een telefonische enquête
B. een vragenlijst in een meisjesblad
C. een vragenlijst via social media
D. een vragenlijst op straat vlak bij een winkelcentrum

Voorbeeld 2

Bekijk de beschrijving van een aantal statistische onderzoeken. Geef per onderzoek aan of het een kwantitatieve of kwalitatieve variabele betreft. Geef ook aan of deze discreet, continu, nominaal of ordinaal is.

1. Er wordt onderzoek gedaan naar het aantal zwerfkatten per stad.
2. Er wordt een enquête opgesteld over wat de top 3 van favoriete muziekartiesten onder achttienjarigen is.
3. Van alle leerlingen in de derde klassen van een middelbare school wordt de lengte gemeten. Een jaar later wordt dit herhaald, om te meten hoeveel iedereen in de tussentijd is gegroeid.
4. Onder kleurenblinde mensen in een bejaardentehuis wordt onderzocht welke mensen wat voor soort kleurenblindheid hebben.
5. De gemiddelde snelheid van mannelijke olympische wedstrijdrenners over de loop van de jaren wordt vergeleken met die van vrouwen.

Antwoord

1. Kwantitatief, discreet: het aantal katten heeft geen tussenwaarden.
2. Kwalitatief, ordinaal: artiesten zijn niet uit te drukken in nummers, maar een top 3 wel in rangordes.
3. Kwantitatief, continu: lengte is een maat met veel mogelijke tussenwaarden.
4. Kwalitatief, nominaal: het soort kleurenblindheid is een naam en niet in een getal uit te drukken.
5. Kwantitatief, continu: snelheid is een maat met veel mogelijke tussenwaarden.

Opgave 7

Bekijk de beschrijving van een aantal statistische onderzoeken. Geef per onderzoek aan of het een kwantitatieve of kwalitatieve variabele betreft. Geef ook aan of deze discreet, continu, nominaal of ordinaal is.

a Met een enquête worden mensen gevraagd of ze wel of niet gelovig zijn, en zo ja, welk geloof ze aanhangen.

b Voor een gepaste opvang op scholen wordt er onderzoek gedaan naar hoeveel kinderen met een autismespectrumstoornis gediagnosticeerd zijn.

c Leerlingen in een basisschoolklas worden gevraagd hun drie favoriete tekenfilms op volgorde op te schrijven.

d De groei van bacterieculturen wordt onder verschillende omstandigheden gemeten aan de hand van het gewicht in petrischaaltjes.

e Aan de hand van het onderzoeken van de verkoop van vliegtickets in vakantieperiodes stelt een maandblad een top 10 van populaire vakantiebestemmingen op.

Voorbeeld 3

Een onderzoeker wil weten of middelbare scholieren tegenwoordig meer kleedgeld dan vroeger krijgen. Onderzoek wat de knelpunten in deze vraag zijn en probeer een formulering te vinden die het onderzoek mogelijk maakt.

Antwoord

Onduidelijk is wat met ‘meer’ wordt bedoeld. Je zou kunnen vragen: hoeveel zakgeld krijgen middelbare scholieren? Een klassenindeling kan handig zijn, bijvoorbeeld € 0,00 - € 5,00 per week, € 5,00 - € 10,00 per week, € 10,00 - € 15,00 per week.

Om welke leerlingen gaat het? Brugklassers krijgen minder dan achttienjarigen. Misschien is het duidelijker als je een leeftijdsklasse neemt, bijvoorbeeld de vijftien- tot zeventienjarigen.

Met welke periode vergelijk je dit? Met de vorige eeuw of met twintig jaar geleden? Bepaal dit en benoem de periode.

De vraag zou nu kunnen zijn: hoeveel geld per week krijgen vijftien- tot zeventienjarigen nu (2017) ten opzichte van zeventien jaar geleden (2000)?

Opgave 8

Formuleer een duidelijke onderzoeksvraag bij de probleemsituaties.

a Hoe betrokken zijn vwo-leerlingen bij de politiek?

b Wat is het succes van bijles?

Verwerken

Opgave 9

Een fabriek produceert autobanden. Omdat een bestelling achterloopt op schema onderzoekt de fabrieksmanager of de dagelijkse productie wel hoog genoeg is. Dan kan hij nadenken over verdere maatregelen.

Deze situatie vraagt om een statistisch onderzoek. Benoem de populatie, de variabele en het soort variabele.

Opgave 10

Stel je voor dat een onderzoeker voor een onderzoek naar de filedruk in Nederland naar waddeneiland Texel gaat en daar 's nachts het aantal auto's op een van de wegen checkt. Hij noteert per voorbijrijdende auto het merk en de snelheid.

a Leg uit waarom dit geen aselect onderzoek is.

b Leg uit waarom dit geen representatief onderzoek is.

c Wat voor soort variabele (kwalitatief, kwantitafief discreet of kwantitatief continu) zijn de volgende variabelen?

- automerk
- aantal auto's per uur
- snelheid

Opgave 11

In de jaren 1982-1988 werd onder 22000 mannelijke Amerikaanse artsen onderzoek gedaan naar de invloed van aspirine op hart- en vaatziekten bij de gemiddelde Amerikaanse man. De helft van de artsen gebruikte om de dag 300 milligram aspirine, wat ongeveer gelijkstaat aan een 'gewoon' aspirientje. De andere helft slikte een placebo ('fopmiddel'). Van de aspirineslikkers kregen 104 personen een hartinfarct, van de placeboslikkers waren dat er 189. De conclusie van het onderzoek was dat het risico op een hartinfarct met ongeveer 45% wordt verlaagd door het slikken van aspirine. Dat dit grote verschil aan toeval was te wijten, vond men uitgesloten vanwege het grote aantal mensen dat aan de studie meewerkte.

a Waarom is hier geen sprake van een representatieve steekproef?
Hoe had deze steekproef moeten worden samengesteld?

b Waarom werd er van placebo's gebruikgemaakt?

c Hoeveel procent van de 11000 aspirineslikkers heeft baat gehad bij het slikken van aspirine?

d Volgens de tekst wordt de kans op een hartinfarct met 45% verlaagd.
Klopt dit?

Opgave 12

In een straat staan precies honderd woningen. Het zijn twintig blokken van vijf woningen. Aan iedere kant van de weg staan tien blokken. Je hebt een even kant met de huisnummers 2 tot en met 100, met een tuin op het zuiden. Je hebt een oneven kant met de huisnummers 1 tot en met 99, met een tuin op het noorden.

a Een energiebedrijf wil het gasverbruik in deze straat onderzoeken. Het bedrijf neemt een steekproef van tien huizen: de huisnummers 1, 11, 21, 31, 41, 51, 61, 71, 81 en 91. Is deze steekproef aselect getrokken?

b Het gemiddelde gasverbruik dat de onderzoeker bij de tien huizen vindt, blijkt veel hoger te zijn dan het gemiddelde in de straat in werkelijkheid blijkt te zijn. Hoe kan dat?

c Bedenk een manier om aselect tien huizen uit de straat voor het onderzoek te selecteren, zodat het gemiddelde gasverbruik van de tien huizen representatief is voor de hele straat.

Opgave 13

De 'Nationale Doorsnee' was in 2000 een landelijk statistiekproject voor leerlingen uit leerjaar 1 en 2. Centrale vraag was: wie is de gemiddelde leerling van Nederland? Het ging bij dit project om negen kenmerken:

- lichaamslengte
- ontbijtgewoonte
- tijdsbesteding sport
- tijdsbesteding tv
- tijdsbesteding computer
- leukste vak op school
- zakgeld per week
- bijverdienste per week
- favoriete popster of popgroep

a Naar welk soort variabele verwijst elk van deze kenmerken? Kwalificeer ze met kwalitatief, kwantitatief, discreet of continu.

b Bedenk bij elk kenmerk een goede vraag die aansluit bij de door jou genoemde soort variabele.

c Welk tiende kenmerk en welke tiende vraag zou je kunnen toevoegen om de gemiddelde leerling van Nederland nog verder te typeren?

Opgave 14

Veel onderzoek gebeurt door mensen een vragenlijst te laten beantwoorden. Het opstellen van de juiste vragen is erg belangrijk. Op slechte vragen krijg je slechte antwoorden. Stel je bent nieuwsgierig wat de leerlingen uit je klas bij het ontbijt eten.

a Je bedenkt als vraag: "Wat vind je lekkerder op de boterham, hagelslag of kaas?" Leg uit waarom deze vraag hier niet goed is.

b Je bedenkt ook de vraag: "Wat is gezonder: een witte boterham of een bruine boterham?" Leg uit waarom ook deze vraag hier niet goed is.

c Je zou ook aan elke leerling kunnen vragen: "Schrijf op wat je vanmorgen hebt gegeten als ontbijt." Wat is een nadeel van deze vraag?

d Je zou kunnen vragen: "Geef met een kruisje aan wat je vanmorgen als ontbijt hebt gehad. Kies uit bruin brood, yoghurt met muesli en/of fruit."

Wat is er mis met deze vraag?

e Welke vraag zou jij stellen waarop je een zinvol antwoord krijgt?
Probeer uit of het een handige en goede vraag is.

Toepassen

Opgave 15: Eigen onderzoek: opzet onderzoek fietsgebruik

Maak een opzet voor een onderzoek onder een deel van de ouders en leerlingen van je school naar hun gebruik van de fiets. Neem je onderzoeksplannen onder de loep en verbeter ze.

a Ontwerp minimaal twee concrete onderzoeksvragen: wat wil je precies weten over het fietsgebruik van de leerlingen en hun ouders? Anders gezegd: welke vragen wil je beantwoord zien nadat je dit onderzoek hebt uitgevoerd?

b Ontwerp een lijst met vragen, zodanig dat je met de antwoorden erop je eigen onderzoeksvragen kunt beantwoorden.

c Stel een lijst met variabelen samen waarin je de antwoorden op je vragenlijst vastlegt en geef bij elke variabele aan of het om een kwalitatieve, discreet kwantitatieve of continu kwantitatieve variabele gaat.

d Ontwerp een aselecte en representatieve steekproef van leerlingen en ouders op je school.

e Ontwerp de manier waarop je de steekproefpersonen de vragen gaat stellen zodanig dat je de meeste kans hebt op zo veel mogelijk serieuze antwoorden.

Testen

Opgave 16

Een wetenschappelijk instituut werkt mee aan het ontwerp van een nieuwe soort satelliet. Hierbij onderzoeken wetenschappers de gemiddelde omlooptijden (de tijdsduur van één baan rond de aarde) van al bestaande satellieten. Met de resultaten kunnen ze het ontwerp van de nieuwe satelliet bevorderen.

Deze situatie vraagt om een statistisch onderzoek. Benoem de populatie, de variabele en het soort variabele.

Opgave 17

In 1954 werd onder 1080860 rijke Amerikaanse kinderen onderzoek gedaan naar de invloed van poliovaccin ‘Salk’ op het wel of niet ontwikkelen van verlammingsverschijnselen. Polio, ook wel kinderverlamming genoemd, is een besmettelijke virusziekte.

De helft van de kinderen kreeg het vaccin ingespoten. De andere helft kreeg een placebo (‘fopmiddel’) ingespoten. Van de daadwerkelijk gevaccineerden kregen 184 kinderen verlammingsverschijnselen, van de placebo-ontvangers waren dat er 497.

a Waarom is hier geen sprake van een representatieve steekproef? Hoe had deze steekproef moeten worden samengesteld?

b Waarom werd er van placebo’s gebruikgemaakt?

c Hoeveel procent van de 540430 daadwerkelijk gevaccineerde kinderen heeft baat gehad bij het Salkvaccin?

d Iemand zegt dat het vaccin de kans op verlammingsverschijnselen met 63% verlaagt. Hoe komt deze persoon daar bij?