5.1 Van steekproef naar populatie

Inleiding

In werkelijkheid is het vaak onmogelijk of te duur om alle te onderzoeken waardes van een toevalsvariabele te achterhalen. Dit geldt ook voor normaal verdeelde toevalsvariabelen zoals 'de lengte van een zonnebloem in midden-Frankrijk in het jaar 2015'. Hoe kun je daar ooit met zekerheid het gemiddelde of de standaardafwijking van kennen? Stuk voor stuk opmeten zal een enorme en prijzige operatie zijn.

Figuur 1

Je leert in dit onderwerp

- hoe je onderzoek doet op basis van goede (aselecte, representatieve en voldoende grote) steekproeven;
- wat de centrale limietstelling is.

Voorkennis

- de normale verdeling met zijn vuistregels over het gemiddelde en de standaardafwijking;
- herkennen of een frequentieverdeling normaal is of niet;
- de wortel-n-wet gebruiken.

Verkennen

Opgave V1

In uitspraken in kranten, boeken en op internet kom je vaak resultaten van statistisch onderzoek tegen. Hier zie je daar een voorbeeld van.

> Uit onderzoek van het Centraal Bureau voor de Statistiek (CBS) blijkt dat bijna de helft van de jongeren tussen de 15 en 25 jaar gebruik maakt van internet op de telefoon. Dat is veel meer dan vorig jaar, toen nog maar 20 procent van de jongeren internette op hun mobiel.
> (Bron: jongeren.blog.nl maart 2010)

Het Centraal Bureau voor de Statistiek (CBS) heeft zich kennelijk afgevraagd hoe het zit met het internetgebruik onder jongeren. Met zo'n probleemstelling begint statistisch onderzoek. De probleemstelling wordt vertaald in een aantal **onderzoeksvragen**. Die vragen worden zo geformuleerd dat de antwoorden data opleveren die statistisch verwerkt kunnen worden om antwoord te geven op het gestelde probleem.

Bekijk de uitspraak hierboven van maart 2010.

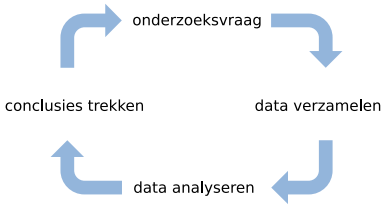
a Welke onderzoeksvraag heeft het CBS zich gesteld?

b Kun je bedenken hoe het CBS dit heeft aangepakt?

c Hoe zou je zelf zo'n onderzoeksvraag aanpakken?

Uitleg 1

Als je kenmerken van een groep mensen of dingen wilt leren kennen, doe je statistisch onderzoek. Daarbij doorloop je in principe altijd de zogenoemde statistische cyclus.



Figuur 2

Vaak is het onmogelijk of te duur om de volledige groep, de populatie, te onderzoeken. Maar op basis van steekproeven uit de populatie kun je ook betrouwbare uitspraken over de gehele populatie doen. Dat heet verklarende statistiek.

Tot nu toe was je vooral bezig met beschrijvende statistiek en dat betreft ‘data verzamelen’ en ‘data analyseren’.

Stel, je wilt weten wat ‘de gemiddelde lengte van een zonnebloem in midden-Frankrijk dit jaar’ is. Hoe kun je daar met zekerheid het gemiddelde of de standaardafwijking van bepalen? Stuk voor stuk opmeten is onmogelijk. Maar met een steekproef kun je een schatting van de gemiddelde zonnebloemlengte maken.

Verklarende statistiek helpt ook bij het onderzoeken naar bijvoorbeeld het verband tussen de zonnebloemlengte en de hoeveelheid regen die in het betreffende jaar in midden-Frankrijk viel. Ook kun je zo de zonnebloemlengte in midden-Frankrijk in het ene jaar vergelijken met die in het andere jaar.

Steeds trek je een steekproef. Die moet wel aselect, representatief en voldoende groot zijn.

Aselect betekent dat elk element van de populatie een even grote kans heeft om in de steekproef te komen.
Representatief betekent dat alle kenmerken die je onderzoekt in de steekproef en in de populatie naar verhouding even vaak voorkomen.
En hoe groter de steekproef, hoe nauwkeuriger de resultaten. In veel gevallen is $n \geq 30$ groot genoeg, maar de minimale steekproefomvang hangt af van wat je onderzoekt.

Opgave 1

De ‘lengte van een zonnebloem in midden-Frankrijk dit jaar’ is, bekeken vanuit een statistisch onderzoek, een toevalsvariabele.

a Geef een beargumenteerde definitie van deze toevalsvariabele en gebruik daarbij onder andere de termen kwantitatief/kwalitatief en wel/niet normaal verdeeld.

b Beschrijf de te onderzoeken populatie.

c Beschrijf een manier om een steekproef uit de door jou beschreven populatie aselect te maken en om de steekproef representatief te maken.

d Leg uit waarom in een betrouwbaar onderzoek een steekproef met een omvang die kleiner is dan 10 zonnebloemen of die bestaat uit alle zonnebloemen in midden-Frankrijk in dit jaar, nooit gebruikt zal worden.

Opgave 2

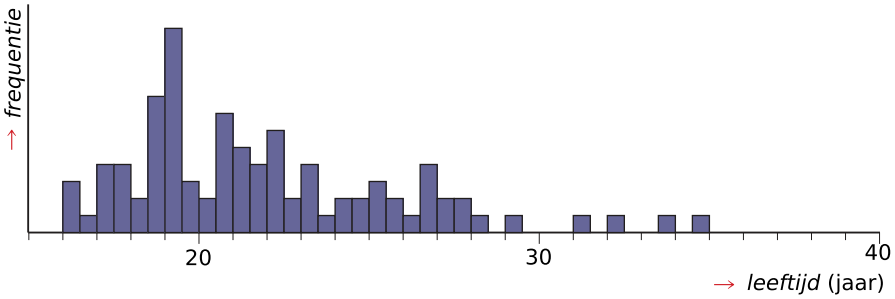
Leg uit of er in de omschreven situaties sprake is van een aselecte en representatieve steekproef.

a Om onderzoek te doen naar het discotheekbezoek onder 14- tot 18-jarigen kies je de leerlingen uit je klas.

b Om uit te zoeken op welke politieke partij Nederlanders stemmen bij de Tweede Kamerverkiezingen, worden uit het bevolkingsregister van Nederland willekeurig 7500 stemgerechtigde inwoners gekozen.

Uitleg 2

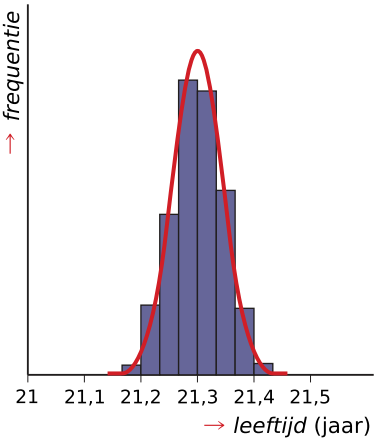
Er is een heel groot concert met tienduizenden bezoekers. De organisatoren van het concert willen de gemiddelde leeftijd van de bezoekers weten.
Bij elk van de 50 ingangen zetten ze een enquêteur die aan elke 10^{e} bezoeker de leeftijd vraagt. Zo worden er 50 steekproeven genomen. Ga ervan uit dat deze steekproeven representatief zijn. Omdat niet iedereen wordt ondervraagd, kun je de gemiddelde leeftijd niet precies te weten komen. Je kunt deze alleen maar schatten. In het linker histogram zijn de gegevens van één van de 50 steekproeven weergegeven.



Figuur 3

Het lijkt erop dat de leeftijden van de concertbezoekers niet normaal verdeeld zijn. Van alle 50 steekproeven die genomen zijn, is de gemiddelde leeftijd berekend, bekijk het rechter histogram. Het lijkt er op dat de steekproevenverdeling wel bij benadering normaal verdeeld is.

In de steekproef gaat het om n onafhankelijke gelijke toevalsvariabelen X. De som S van deze gelijke toevalsvariabelen is bij benadering normaal verdeeld met gemiddelde $\overline{S} = n \cdot \overline{X}$ en standaardafwijking $\sigma(S) = \sqrt{n} \cdot \sigma(X)$. Ook het gemiddelde van S is bij benadering normaal verdeeld met een gemiddelde van $\overline{S} = \overline{X}$ en een standaardafwijking van $\sigma\left(\overline{S}\right) = \frac{\sigma(X)}{\sqrt{n}}$.



Figuur 4

Hoe meer steekproeven je doet, hoe beter de benadering is. In veel gevallen is $n \geq 30$ groot genoeg, maar de minimale steekproefomvang hangt af van wat je onderzoekt.

In het algemeen zijn de uitkomsten van een grote hoeveelheid onafhankelijke steekproeven uit dezelfde populatie bij benadering normaal verdeeld. Hoe groter het aantal steekproeven, hoe beter de benadering. Hieruit volgt dat de steekproevenverdeling bij benadering normaal verdeeld is. Dit wordt de centrale limietstelling genoemd.

Opgave 3

Een vulmachine vult pakken suiker. Het gemiddelde gewicht van een pak suiker is 1000 gram en de standaardafwijking 2,8 gram.
De Consumentenbond doet een steekproef van honderd pakken.

a Waarom mag je zeggen dat het totale gewicht van die honderd pakken bij benadering normaal verdeeld is? Is het daarbij belangrijk dat het gewicht van een pak suiker normaal verdeeld is?

b Is de steekproevenverdeling ook bij benadering normaal verdeeld?

c Wat is het gemiddelde en de standaardafwijking van de steekproevenverdeling?

d De Consumentenbond vindt het onacceptabel als het gemiddelde gewicht van de pakken suiker uit de steekproef kleiner is dan 999 gram.
Bereken in vier decimalen de kans dat dit het geval is.

Opgave 4

Twee onderzoekers geven elk een werkwijze die beide een bij benadering normale kansverdeling opleveren.

Onderzoeker A neemt een steekproef van 5000 zonnebloemen uit midden-Frankrijk en meet de lengte ervan. Hij maakt er een histogram van.

Onderzoeker B neemt duizend steekproeven van ieder 5000 zonnebloemen uit midden-Frankrijk en meet de lengte van al deze zonnebloemen. Van ieder van de 1000 steekproeven berekent ze de gemiddelde zonnebloemlengte. Van de 5000 gemiddelde zonnebloemlengtes maakt zij een histogram.

a Beschrijf voor beide werkwijzen van welke toevalsvariabele de onderzoekers een histogram maken.

b Geef aan welke van de twee werkwijzen vanwege de centrale limietstelling een normale kansverdeling oplevert.

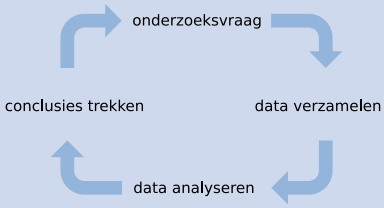
c Een groep scholieren wil ook aan de slag met zonnebloemlengtes.

Moeten de scholieren voor hun onderzoek meerdere steekproeven trekken of slechts één (die natuurlijk wel aselect, representatief en groot genoeg is) om toch een betrouwbare uitspraak over alle zonnebloemen te kunnen doen?

Theorie en voorbeelden

Om te onthouden

Een statistisch onderzoek doorloopt in principe altijd de zogenoemde **statistische cyclus**.



Figuur 5

Data verzamelen van de volledige populatie die je wilt onderzoeken, is vaak erg duur en soms ook onmogelijk. Gelukkig is het meestal wel mogelijk om een **aselecte**, **representatieve steekproef** van voldoende omvang uit de **populatie** te trekken en op basis daarvan betrouwbare aannames te doen over de volledige populatie. Deze tak van wetenschap heet **verklarende statistiek**.

Een belangrijke stelling die bij verklarende statistiek wordt gebruikt is de **centrale limietstelling**: De som van een groot aantal onafhankelijke, mogelijk verschillende, willekeurig verdeelde toevalsvariabelen is bij benadering normaal verdeeld. De toevalsvariabelen zelf hoeven niet normaal verdeeld te zijn.

De kansverdeling van de gemiddelde steekproefuitslagen heet de **steekproevenverdeling**. Daarbij gaat het over n onafhankelijke gelijke toevalsvariabelen X. De som S van deze gelijke toevalsvariabelen is bij benadering normaal verdeeld met gemiddelde $\overline{S} = n \cdot \overline{X}$ en standaardafwijking $\sigma(S) = \sqrt{n} \cdot \sigma(X)$. (Denk aan de wortel-n-wet.)

Ook het gemiddelde van S is bij benadering normaal verdeeld met een gemiddelde van $\overline{S} = \overline{X}$ en een standaardafwijking van $\sigma\left(\overline{S}\right) = \frac{\sigma(X)}{\sqrt{n}}$.

Wel moet n voldoende groot zijn. Wat voldoende groot is om de centrale limietstelling te gebruiken, is afhankelijk van de verdeling van de toevalsvariabelen. In veel gevallen is $n \geq 30$ groot genoeg.

De centrale limietstelling wordt vaak gebruikt bij het doen van steekproeven. Je kunt dan uitspraken doen over een populatie zonder de hele populatie apart te onderzoeken.

Behalve uit het achterhalen van de waarde van een populatiekenmerk bestaat verklarende statistiek ook uit onderzoek naar het verband tussen meerdere populatiekenmerken en naar overeenkomst/verschil tussen meerdere populaties.

Voorbeeld 1

Zijn de volgende steekproeven representatief? Geef ook aan of de steekproef aselect is.

1. Er wordt onderzoek gedaan naar het stemgedrag voor politieke partijen van Nederlanders. Daarvoor worden aselect achtduizend stemgerechtigde Nederlanders uit het bevolkingsregister gekozen.
2. Er wordt onderzoek gedaan naar de hoeveelheid verf in blikken. Van alle geproduceerde blikken wordt steeds het tiende blik gewogen.
3. Er wordt onderzoek gedaan naar het aantal uren dat ouders van de kinderen van een school van huis zijn. De ouders die op een ouderavond komen, worden ondervraagd.

Antwoord

1. De steekproef is, behalve aselect, waarschijnlijk ook representatief door de grote omvang.
2. De steekproef is waarschijnlijk wel representatief, want de steekproef is uit de hele populatie. De steekproefomvang is goed en meestal maakt het niet uit van welke blikken je de hoeveelheid meet. (Het zou natuurlijk kunnen dat elk tiende blik door één persoon wordt gevuld en de rest door een ander.) De steekproef is niet aselect, want niet ieder blik heeft een even grote kans om in de steekproef terecht te komen.
3. De steekproef is niet representatief, want waarschijnlijk komen er naar verhouding minder ouders die veel van huis zijn naar de ouderavond. De steekproef is ook niet aselect, want ouders die niet op een ouderavond komen, hebben geen kans om in de steekproef terecht te komen.

Opgave 5

Zijn de volgende steekproeven aselect? Licht je antwoord toe.

a Een onderzoek onder leerlingen van twintig scholen, waarbij van elke school 10% van de leerlingen willekeurig wordt gekozen.

b Een onderzoek dat gebruikmaakt van vragenlijsten die zijn ingevuld door willekeurig gekozen mensen van 65 jaar en ouder.

c Een onderzoek naar de tevredenheid van abonnees van een telefoonbedrijf door willekeurig gekozen klanten die de helpdesk bellen te ondervragen.

Opgave 6

Geef van de volgende steekproeven aan of ze representatief en/of aselect zijn. Licht je antwoord toe.

a Een leverancier van koffie doet onderzoek naar de tevredenheid van zijn klanten over de koffie. Hij ondervraagt alle klanten die in zijn winkel komen.

b De RDW is de organisatie die beslist welke voertuigen in Nederland worden toegelaten op de wegen. De RDW doet onderzoek naar voertuigen. De RDW neemt daarom willekeurig een grote steekproef uit de lijst van alle kentekens van voertuigen en stuurt de eigenaars van de voertuigen een vragenlijst over winterbanden.

c Een onderzoeksbureau doet onderzoek naar gebruik van sociale media onder jongeren. Het bedrijf verspreidt een enquête via sociale media.

Voorbeeld 2

Een machine vult pakken suiker. Het gewicht van een pak suiker is 501 gram met een standaardafwijking van 2,9 gram.

Er wordt een steekproef genomen van 50 pakken.

Hoe groot is de kans dat het gemiddelde gewicht van een pak suiker uit deze steekproef minder is dan 500 gram?

Antwoord

Vanwege de centrale limietstelling mag je ervan uitgaan dat het gemiddelde gewicht $\overline{X}$ van de steekproevenverdeling normaal verdeeld is waarbij het gemiddelde 501 gram is en de standaardafwijking $\frac{2,9}{\sqrt{50}}$ gram.

De gevraagde kans is $\mathrm{P}\left(\overline{X} < 500 | \mu = 501 \text{ en } \sigma = \frac{2,9}{\sqrt{50}}\right) \approx 0,007$.

Dit bereken je met de grafische rekenmachine.

Opgave 7

Gebruik de gegevens uit **Voorbeeld 2**.

a Moet het gewicht van een pak suiker per se normaal verdeeld zijn?

b Reken na dat de kans ongeveer 0,007 is.

c Als de steekproefgrootte 100 was geweest, hoe groot was dan in vier decimalen de kans geweest?

Opgave 8

In een wetenschappelijke analyse door het blad ‘Science’ (gepubliceerd in maart 2016) van allerlei onderzoeken op het gebied van de experimentele economie is ook een onderzoek uit 2011 bestudeerd dat als resultaat gaf: ‘ruilen is minder effectief dan betalen’.

Tijdens de analyse werd geconcludeerd dat voor dit onderzoek het aantal proefpersonen veel te klein was. Bij een herhaling van dit onderzoek kan het effect niet worden gereproduceerd.

Verklaar met statistische termen en argumenten dat het kleine aantal proefpersonen en het feit dat het effect niet reproduceerbaar is, met elkaar samenhangt.

Denk aan termen als populatie, steekproef, steekproefomvang, steekproefgemiddelde, normale verdeling, wortel-n-wet, centrale limietstelling.

Verwerken

Opgave 9

In de Nationale Wetenschapsquiz kwam de volgende vraag voor. Stel, je wilt weten hoeveel schoolgaande kinderen er gemiddeld per gezin zijn. Je neemt een grote steekproef onder schoolkinderen en vraagt hun hoeveel schoolgaande broers en zussen zij hebben. Op basis daarvan bepaal je het gemiddelde aantal schoolgaande kinderen per gezin.

Is dit een goede aanpak? Welk van de antwoorden is correct en waarom?

A. Ja, zo krijg je een juiste schatting.
B. Nee, zo krijg je een te lage schatting.
C. Nee, zo krijg je een te hoge schatting.

Opgave 10

Een machine vult pakken suiker. Het gemiddelde gewicht van een pak suiker is 1002 gram met een standaardafwijking van 6,5 gram.
Er wordt een steekproef genomen van 100 pakken.

a Waarom mag je ervan uitgaan dat de steekproevenverdeling normaal verdeeld is?

b Bereken in vier decimalen de kans dat het gemiddelde gewicht van een pak suiker uit de steekproef minder is dan 1000 gram.

Opgave 11

Er wordt onderzoek gedaan naar het aantal Nederlanders dat Fries spreekt. Er wordt een steekproef genomen van 1200 Nederlanders: uit elke provincie aselect 100 inwoners.

a Lever commentaar op deze steekproefsamenstelling.

b Hoe kan het beter? Geef hiervoor minstens twee manieren.

Opgave 12

Leg uit of bij de volgende situaties uitgegaan mag worden van normale verdeling.

a Het aantal ogen dat je gooit als je drie keer met een dobbelsteen gooit.

b Het aantal keer kop als je 10000 keer met een geldstuk werpt.

c De gemiddelde score van een boogschutter die honderd keer achter elkaar op een schietschijf schiet met daarop de scores 1 tot en met 10.

Opgave 13

Een onderzoeker wil de gemiddelde lengte van een zonnebloem in midden-Frankrijk weten in het jaar 2016. Daarvoor heeft hij de lengte van duizend zonnebloemen gemeten. De gegevens heeft hij in een relatieve frequentietabel gezet.

lengte zonnebloem (meter)	percentage
0– < 0,5	0,1
0,5– < 1	1,1
1– < 1,5	2,6
1,5– < 2	9,1
2– < 2,5	21,6
2,5– < 3	31,4
3– < 3,5	19,9
3,5– < 4	9,8
4– < 4,5	3,1
4,5– < 5	1,0
5– < 5,5	0,3

Tabel 1

a Bereken het steekproefgemiddelde van de getoonde steekproef van 1000 zonnebloemen. Bereken ook de bijbehorende standaardafwijking. Geef beide waardes in meter en rond af op twee decimalen.

b De steekproef is voldoende groot om de centrale limietstelling te mogen gebruiken. Waarom mag je nu niet zomaar zeggen dat de lengte van een zonnebloem normaal verdeeld is?

Neem aan dat de lengte van een zonnebloem in midden-Frankrijk in 2016 normaal verdeeld is met een gemiddelde van 2,83 en een standaardafwijking van 76 cm.

c Bereken in vier decimalen de kans dat een zonnebloem uit midden-Frankrijk in 2016 een lengte heeft die kleiner is dan het steekproefgemiddelde van de 1000 zonnebloemen.

Opgave 14

Een steekproef van 30 of groter is in veel gevallen voldoende groot om de centrale limietstelling te mogen toepassen. Soms is een kleinere steekproef al voldoende, maar soms moet de steekproef ook een stuk groter zijn dan 30.

Stel, er zijn twee dobbelstenen. Een gewone en een bijzondere dobbelsteen met de waarden 1, 2, 4, 5, 6, 6. Je gooit beide dobbelstenen een groot aantal keer en berekent het gemiddeld aantal ogen.

a Bij welke dobbelsteen, denk je, zal het gemiddeld aantal ogen dat je gooit eerder een normale verdeling benaderen?

b Is de som van de gemiddeldes van beide dobbelstenen (bij benadering) normaal verdeeld, als je maar vaak genoeg gooit?

Toepassen

Opgave 15: Populaire webwinkel

Een Britse populaire webwinkel heeft gedurende Britse kantooruren gemiddeld 65000 bezoekers per uur, met een standaarddeviatie van 27500 bezoekers: het aantal bezoekers is normaal verdeeld. Tijdens een test wordt een aselecte steekproef van 50 kantooruren getrokken en wordt steekproefgemiddelde $\overline{B}$ van het aantal websitebezoekers berekend.

a Welke kansverdeling heeft $\overline{B}$? Beargumenteer je antwoord.

b Hoe groot is de kans dat het steekproefgemiddelde van deze 50 kantooruren hoger is dan 73000 bezoekers per kantooruur? Rond af op vier decimalen.

c Wat is waarschijnlijker: dat het steekproefgemiddelde $\overline{B}$ lager ligt dan 60000 bezoekers per kantooruur of dat er tijdens een willekeurig kantooruur minder dan 60000 bezoekers zijn? Beargumenteer je antwoord.

d Leg uit waarom het antwoord op de vorige vraag overeenkomt met wat je volgens de centrale limietstelling van een steekproef mag verwachten.

Testen

Opgave 16

Leg uit of de gekozen steekproef groot genoeg is.

a In het buitenland komt een ziekte bij 0,01% van de mensen voor. De overheid wil weten of de ziekte ook in ons land voorkomt en onderzoekt daarom vijfduizend Nederlanders.

b De manager van een ijsbaan wil weten of er op woensdag meer meisjes dan jongens komen schaatsen. Op een zeker moment neemt hij een steekproef van dertig bezoekers en telt het aantal meisjes.

Opgave 17

Zijn bij de volgende onderzoeken de steekproeven aselect en representatief? Leg je antwoord uit.

a Een onderzoek onder treinreizigers, waarbij de vragenlijsten willekeurig worden verstuurd naar bezitters van een OV-chipkaart. Ga ervan uit dat iedere reiziger maar één OV-chipkaart bezit.

b Een onderzoek houden of Nederlanders voldoen aan de beweegnorm door een maand lang alle bezoekers van de website van de Gezondheidsraad te bevragen.

Opgave 18

Bij een snackbar kan iedere klant zelf een portie mayonaise toevoegen met een mayonaisepomp. De mayonaisepomp pompt per portie gemiddeld 25 gram mayonaise op met een standaardafwijking van 3 gram.

Bij een steekproef van 40 porties blijkt dat het steekproefgemiddelde 18 gram is.

a Bereken de kans op een steekproefgemiddelde van minder dan 18 gram.

b Wat vind je van de waarschijnlijkheid van genoemd steekproefgemiddelde als het echt waar is dat het gemiddelde portievolume gelijk is aan 25 gram met een standaardafwijking van 3 gram?