Populaties beschrijven met kansmodellen

Prof. dr. Herman Callaert

Deze tekst probeert, met voorbeelden, inzicht te geven in de manier waarop je in de statistiek populaties bestudeert. Dat doe je met kansmodellen.
Dit inzicht is cruciaal bij de studie van de normale verdeling. Daar werk je immers met kansmodellen om eigenschappen van normaal verdeelde populaties te achterhalen.

Inhoudstafel

1 Een echte populatie: hoe bestudeer je die?

Als voorbeeld werken we met 2 populaties die elk een eigen type vertegenwoordigen:
- een populatie met discrete uitkomsten (een dobbelsteen)
- een populatie met continue uitkomsten (de lengte).

1.1 Een echte dobbelsteen uit Las Vegas

De dobbelsteen, waarvan je hiernaast een afbeelding ziet, is je "echte" populatie.

Wat is de kans dat "deze" dobbelsteen valt op een 1 of op een 2 of …?

Probleem: het gedrag van de "echte" populatie ken je niet.
Oplossing: werk met een model dat je wel kent.
Vraag: hoe doe je dat?

1.2 De lengte van 17–jarige Vlaamse meisjes, vandaag

De lengte, vandaag, van alle 17–jarige Vlaamse meisjes is de "echte" populatie die je wil bestuderen.

Hoeveel percent van die meisjes is kleiner dan 150 cm, of groter dan 180 cm, of …?

Probleem: het gedrag van de "echte" populatie ken je niet.
Oplossing: werk met een model dat je wel kent.
Vraag: hoe doe je dat?

2 Van een echte populatie naar een theoretisch model

"*In statistiek is abstract redeneren onlosmakelijk verbonden met interpretatie van data en context*" en dus ga je op zoek naar een zinvol theoretisch model vanuit:

- de context van de onderzoeksvraag
  - waarover gaat het (wie, wat, waar, wanneer,...)?
  - wat weet je uit analoge onderzoeken (vroegere studies, literatuurgegevens,…)?
- de data van een steekproef uit de bestudeerde populatie
  - wat vertelt het gevonden cijfermateriaal?
  - welk kansmodel past bij die data?

2.1 Een model voor de dobbelsteen uit Las Vegas

2.1.1 De context

De bestudeerde dobbelsteen lijkt goed op andere dobbelstenen. Uit de symmetrie vermoed je dat elk zijvlak evenwaardig is, of toch ongeveer. Als je goed kijkt, dan zie je wel kleine verschillen. Er is maar één zijvlak waar de tekst "*Las Vegas*" op staat. Ook is het aantal ogen verschillend per zijvlak en die ogen zijn gevormd door lichte uithollingen die wit geverfd zijn. Maar toch denk je dat die verschillen zo klein zijn dat ze bijna geen invloed hebben. En dan kan je misschien te werk gaan zoals ook vroeger veel verschillende mensen op veel verschillende plaatsen bij veel verschillende dobbelstenen hebben gedaan: behandel elk zijvlak evenwaardig (de uniforme verdeling).

De context vertelt je al heel wat maar het gaat toch nog altijd over deze specifieke dobbelsteen uit Las Vegas. Hoe weet je dat die niet op een of andere manier getrukeerd is? En zelfs als hij niet getrukeerd is, gedraagt hij zich dan zoals verwacht?
Om hierover iets te weten, kijk je naar een steekproef, gegenereerd door <u>*deze*</u> dobbelsteen.

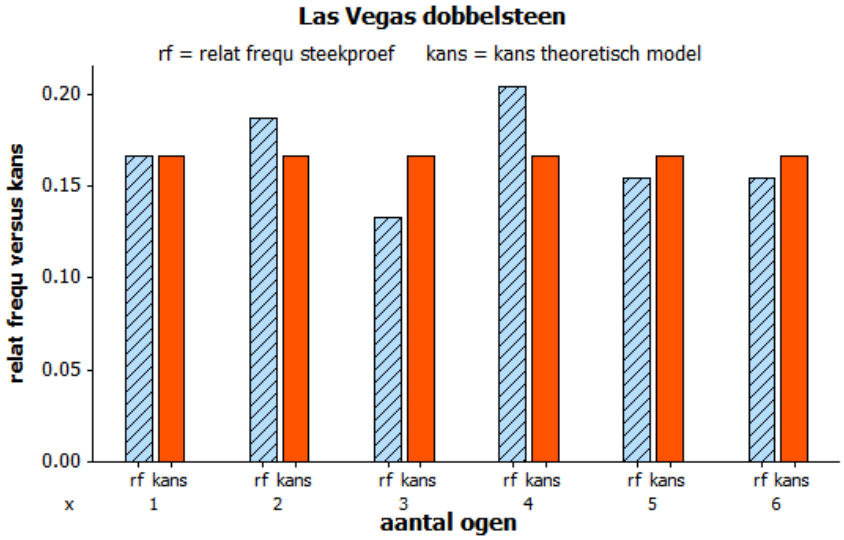
2.1.2 De data

240 worpen van de dobbelsteen uit Las Vegas leverde volgend resultaat:

steekproef $n = 240$				theoretisch model X
aantal ogen x_i	frequentie f_i	relat frequ rf_i		kans $P(X = x)$
1	40	0.1667		0.1667
2	45	0.1875		0.1667
3	32	0.1333		0.1667
4	49	0.2042		0.1667
5	37	0.1542		0.1667
6	37	0.1542		0.1667

In de tabel die de resultaten van de steekproef beschrijft, staan de verschillende uitkomsten samen met de frequentie en de relatieve frequentie. Daarnaast staat, ter vergelijking, een kolom die voor elke uitkomst de kans aangeeft dat het theoretisch model X (= de "ideale" dobbelsteen) één van de mogelijke waarden aanneemt.

De informatie in de tabel kan je ook grafisch in een staafdiagram voorstellen:



Wat vertellen deze data?

Er is de concrete dobbelsteen uit Las Vegas (de concrete populatie) en er is een abstract theoretisch model (dat onderstelt dat alle uitkomsten dezelfde kans hebben). Kan je, op basis van de data, dit model gebruiken om eigenschappen van de dobbelsteen te bestuderen?

Je weet dat uitkomsten van een steekproef aan het toeval onderhevig zijn. Zelfs bij een "perfecte" dobbelsteen heb je andere resultaten als je die dobbelsteen vandaag 240 keer opgooit en morgen nog een keer. De vraag is dus niet: "is er een verschil tussen de steekproef en het model?" Het antwoord hierop is altijd "ja". De echte vraag is: "is het verschil tussen wat je ziet in de steekproef en wat je verwacht volgens het model zo groot dat je besluit om het model niet te gebruiken?".

Als het verschil "niet te groot" is, dan vertrouw je het model. Je gaat er dan vanuit dat de eigenschappen die je met dat model kan berekenen een goede weergave zijn van de eigenschappen van de echte populatie. De kansrekening zegt bijvoorbeeld dat een "ideale" dobbelsteen met kans 2/6 minstens vijf ogen oplevert. En nu zeg je dat dit ook waar is voor die dobbelsteen uit Las Vegas.

Nota.
In de bovenstaande redenering verwerp je het voorgestelde model als "de verschillen te groot zijn". Maar wat is "te groot"? Hiervoor heb je criteria nodig en statistische technieken die (voor de meeste leerlingen) het niveau van het secundair onderwijs overstijgen.
Zo vind je bijvoorbeeld met een "chi-kwadraat toets voor aanpassing" dat er geen significant verschil is tussen de steekproefresultaten van de dobbelsteen en de verwachte waarden volgens de uniforme verdeling (p-waarde = 0.45).

Besluit

Je mag de uniforme verdeling vertrouwen als model voor de dobbelsteen uit Las Vegas.

2.2 Een model voor de lengte van 17–jarige Vlaamse meisjes

2.2.1 De context

De lichaamslengte is een grootheid die al jaar en dag voor verschillende doeleinden is bestudeerd. Je weet dat 100 jaar geleden de mensen kleiner waren dan nu, en dat Nederlanders groter zijn dan Chinezen. Maar ondanks al die verschillen in plaats en tijd is toch gebleken dat, qua vorm, de lengte van mensen goed kan beschreven worden door een normale curve.

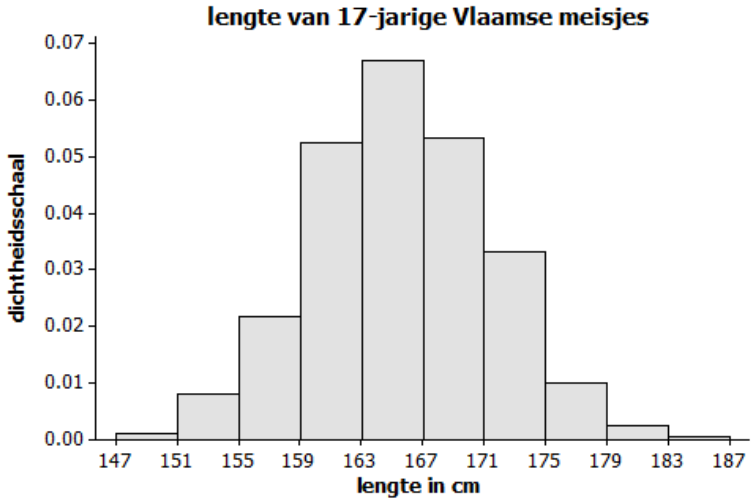
Voor de huidige onderzoeksvraag bestaat de populatie uit lichaamslengtes en de context wijst in de richting van de normale verdeling. Maar is dat ook een goed model voor de concrete populatie van lengtes van alle huidige 17–jarige Vlaamse meisjes? Om dit te weten moet je ook naar de data kijken.

2.2.2 De data

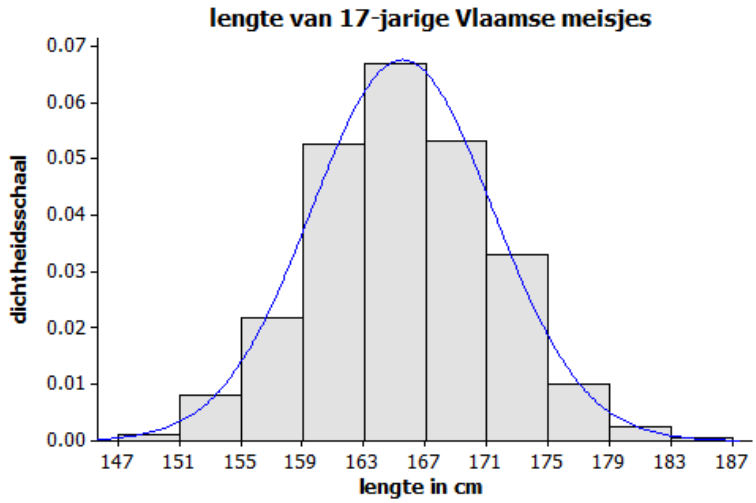
Bij een lukrake steekproef van 400 Vlaamse meisjes van 17 jaar meet je nauwkeurig de lengte. Dat levert 400 getallen met de volgende eigenschap (alles in cm):

minimum	$\min = 147$
maximum	$\max = 185$
gemiddelde	$\bar{x} = 165.5$
mediaan	$Me = 165.7$
standaardafwijking	$s = 5.91$

Je kan deze data ook grafisch voorstellen in een histogram.



Door een histogram op de dichtheidsschaal te tekenen is de totale oppervlakte van het histogram gelijk aan 1. Je weet dat ook de totale oppervlakte onder de normale curve gelijk is aan 1. Je kan dus beide figuren op eenzelfde grafiek tekenen om ze met elkaar te vergelijken.



Juist zoals bij de steekproefresultaten van de Las Vegas dobbelsteen heb je ook hier een verschil tussen de steekproefresultaten van de lengtes en het "ideale" model (de normale). Je moet hierbij **verschillen in oppervlakte** bekijken! [zie de tekst: "Niet de hoogte, wel de oppervlakte" bij de "Werkteksten" op http://www.uhasselt.be/lesmateriaal-statistiek]

Ook hier is de vraag: is het verschil tussen de waarnemingen en het voorgestelde normale model "te groot"? In dat geval kan je dit normale model niet gebruiken om die lengtes te beschrijven.

Om te weten wat "te groot" is heb je terug statistische technieken nodig die (voor de meeste leerlingen) het niveau van het secundair onderwijs overstijgen. Je zou grafisch kunnen starten met een "normal probability plot". Als je verder zou gaan (met formele toetsen) dan zal je vinden dat de waargenomen verschillen hier niet significant zijn.

Besluit

Je mag de normale verdeling vertrouwen als model om de lengtes van 17–jarige Vlaamse meisjes te beschrijven. Als je de populatie van al die lengtes voorstelt door X dan betekent dit dat je werkt met $X \sim N(\mu;\sigma)$. Hierbij is μ het gemiddelde van de populatie en σ is de standaardafwijking van de populatie.

Dat gemiddelde en die standaardafwijking ken je niet en dus schat je hun waarde uit de data van jouw steekproef. Je gebruikt hierbij zowel je berekeningen als je gezond verstand (in de context van het onderzoek). Als je bijvoorbeeld in een of andere steekproef een gemiddelde lengte van $\bar{x} = 165.48271$ zou vinden, dan zal het in de meeste situaties zinvoller en eenvoudiger zijn om te werken met een populatiemodel waarbij je $\mu = 165.5$ neemt.

In het huidige voorbeeld kan je als populatiegemiddelde werken met $\mu = 165.5$ en voor de standaardafwijking van de populatie kan je werken met $\sigma = 5.91$.

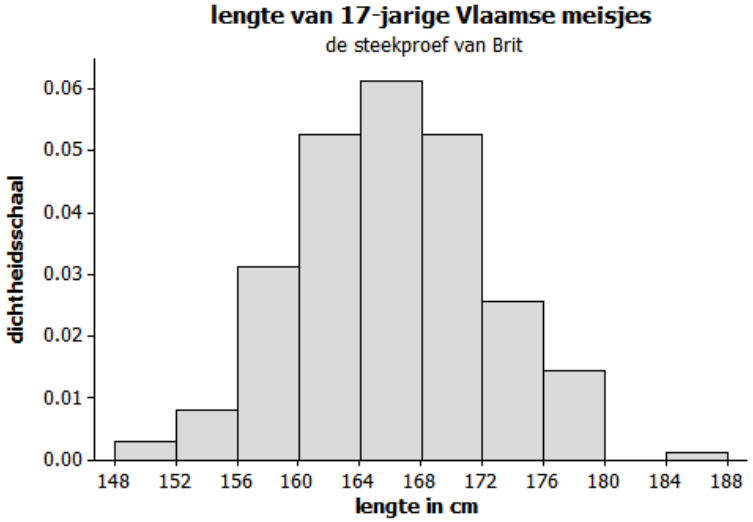
Vanaf nu vertrouw je erop dat je met het kansmodel $X \sim N(165.5;5.91)$ de populatie van de lengtes van 17–jarige Vlaamse meisjes kan bestuderen.

3 Het theoretisch model: notatie en interpretatie

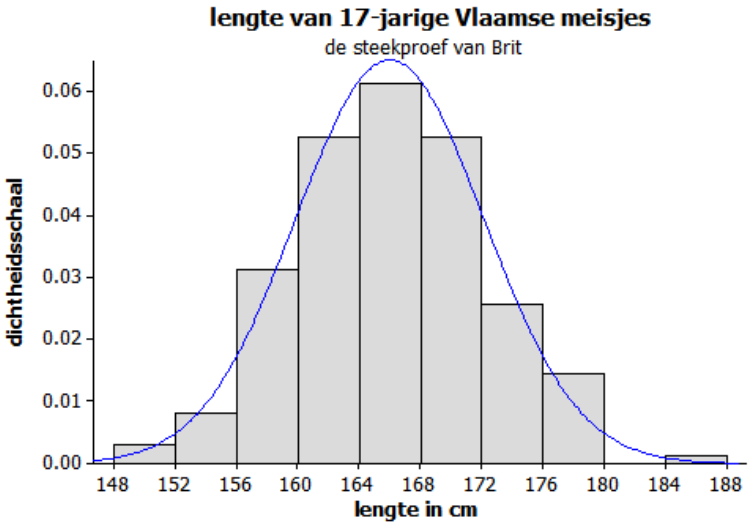
3.1 De lengte van 17–jarige Vlaamse meisjes, deel 2

Hierboven heb je een model gebouwd voor de lengte van 17–jarige Vlaamse meisjes. Iemand anders (het was Brit) heeft dat ook gedaan. Zij deed het op juist dezelfde manier zoals jij en vond voor haar steekproef van 400 lengtes (alles in cm):

minimum	$\min = 149$
maximum	$\max = 185$
gemiddelde	$\bar{x} = 166$
mediaan	$Me = 166.1$
standaardafwijking	$s = 6.14$



Op de figuur lijkt het histogram niet te ver af te wijken van een normale curve (en ook formele statistische toetsen bevestigen dit).



Uit de context (de lichaamslengte) en uit de data (de steekproef van Brit) kan je onderstellen dat je mag werken met een kansmodel dat qua vorm normaal verdeeld is.

Brit besluit dat zij het model $X \sim N(166;6.14)$ kan gebruiken om de lengte van 17–jarige Vlaamse meisjes te bestuderen.

3.2 Onderzoeksvraag, context, data en notatie

De onderzoeksvraag gaat over de beschrijving van de **populatie** van de lengtes van alle Vlaamse meisjes die vandaag 17 jaar zijn.

Uit de context en de data blijkt dat, qua vorm, een normale verdeling een goed kansmodel kan zijn om deze populatie van lengtes te beschrijven. Je hebt dan nog 2 parameters nodig: het gemiddelde en de standaardafwijking.

Populatieparameters stel je voor met Griekse letters en dus schrijf je:

μ: het populatiegemiddelde
= het gemiddelde van de lengtes van alle 17–jarige Vlaamse meisjes

σ: de standaardafwijking van de populatie
= de standaardafwijking van de lengtes van alle 17–jarige Vlaamse meisjes

Zowel jij als Brit besluiten dat $X \sim N(\mu;\sigma)$ een goed **populatiemodel** is, maar wat is μ en σ?

- Jij zegt: mijn beste benadering voor het populatiegemiddelde μ is het gemiddelde dat ik in mijn steekproef heb gevonden ($\bar{x} = 165.5$) en dus werk ik met $\mu = 165.5$. Voor de standaardafwijking vond ik in mijn steekproef $s = 5.91$ en dus neem ik die waarde als mijn beste benadering voor de standaardafwijking van de populatie en werk ik met $\sigma = 5.91$. Alles samen werk ik dus met het model $X \sim N(165.5;5.91)$ om **de populatie** van die lichaamslengtes te beschrijven.

- Brit zegt: ik neem als benadering voor het populatiegemiddelde $\mu = 166$ want in mijn steekproef vond ik $\bar{x} = 166$. Voor de standaardafwijking van de populatie neem ik $\sigma = 6.14$ want de enige informatie die ik heb komt uit mijn steekproef en daar was $s = 6.14$. Ik neem dus $X \sim N(\mu;\sigma)$ met $\mu = 166$ en $\sigma = 6.14$ om **de populatie** van die lichaamslengtes te beschrijven.

3.3 Alle modellen zijn fout maar sommige zijn uitermate nuttig

Met een wiskundig model vang je de werkelijkheid niet… maar is dat ook nodig?

Je mag er zeker van zijn dat de dobbelsteen uit Las Vegas niet "perfect eerlijk" is. Geen enkele dobbelsteen is dat. Maar zolang de afwijkingen minimaal zijn (in het kader van de onderzoeksvraag) is het handig om een wiskundig model (de uniforme verdeling) te gebruiken. Je kan dan vooraf allerlei kansen berekenen als je met die dobbelsteen gaat gooien. Hoe zou je zonder model kunnen weten wat bijvoorbeeld de kans is dat je minstens acht keer moet gooien vooraleer je de eerste zes ziet verschijnen?

Om de lichaamslengte van alle Vlaamse meisjes die vandaag 17 jaar zijn te beschrijven, werk je met een model dat niet perfect is maar meestal toch zeer goed werkt.

- Je gebruikt een wiskundig model (de normale verdeling) waarbij je een zo goed mogelijke schatting maakt van het gemiddelde en de standaardafwijking van de populatie. Als jij in je steekproef $\bar{x} = 165.5$ en $s = 5.91$ vond, dan neem je $\mu = 165.5$ en $\sigma = 5.91$ en werk je met $X \sim N(165.5;5.91)$. Maar als je alleen de informatie van Brit hebt, dan neem je $X \sim N(166;6.14)$ als model voor diezelfde populatie.

- Zelfs al zou je de echte μ en σ van de populatie kennen (hoe zou je die opmeten?) dan nog is het normale model slechts een benadering als je het puur wiskundig bekijkt. De normale dichtheidsfunctie is immers gedefinieerd op $]-\infty,+\infty[$ en is bovendien overal strikt positief. Dat betekent dat $P(-\infty < X < 0)$ niet gelijk is aan nul, terwijl meisjes met een negatieve lengte niet bestaan. Maar in de praktijk werkt het normale model prima!