Position paper: Statistiek in het Voortgezet Onderwijs

Peter Kop[1], Marcel van Assen[2], Ronald Meester[3], Casper Albers[4], Lonneke Boels[5], Marianne van Dijke-Droogers[6] , Jos Tolboom[7], Piet Blokland[8], Lidy Wesker-Elzinga[9], Arjen de Vetten[10], Petra Hendrikse[11], Merijn Smit[12]

De auteurs zijn allen lid van de werkgroep Statistiek van de Nederlandse Vereniging van Wiskundeleraren en geven hier hun visie op hoe statistiekonderwijs en de wetenschappelijke onderzoekscyclus het beste vormgegeven kunnen worden in de bovenbouw van het voortgezet onderwijs.

1 Inleiding

We leven in een complexe wereld en het is menselijk en wenselijk om hier enige grip op te willen krijgen. De discipline statistiek is hierbij een uitstekend hulpmiddel.

Statistiek is de tak van wetenschap die zich bezighoudt met gegevens ('data'). Wat is de meest zinvolle manier om zulke gegevens te verzamelen, presenteren, analyseren, en hoe interpreteren we de uitkomsten van de analyses? Doordat dit fundamenteel verschillende aspecten zijn, is statistiek een veelzijdig, en daarmee ingewikkeld, vakgebied. Kenmerkend voor statistiek is de rol van onzekerheid: welke gegevens verzameld en geanalyseerd worden, en daarmee ook de interpretatie van de analyses, hangen voor een deel van het toeval af. Daarmee is statistiek inherent verbonden aan de kansrekening en andere aspecten van de wiskunde. Door deze onzekerheid onderscheidt statistiek zich van andere wiskundige disciplines waar het vaak gaat om zekerheden. Daarnaast is het in de statistiek essentieel de context – en dus ook andere vakgebieden – erbij betrokken te houden. Het is statistisch goed te zeggen dat roken samenhangt met het krijgen van longkanker, maar voor de verklaring waarom die samenhang er is, is kennis nodig van bijvoorbeeld biologie, scheikunde en geneeskunde.

Statistiek kun je zien als de oudste data science, en veel van wat we hier bespreken is in die zin klassiek. Maar statistiek ontwikkelt zich ook: kunstmatige intelligentie en machine learning zijn in de basis statistische disciplines, die echter zonder kennis van de meer klassieke statistiek niet te begrijpen zijn.

In het voortgezet onderwijs gaat het bij statistiek om het verzamelen van een beperkte hoeveelheid gegevens om bepaalde conclusies te kunnen trekken, of om het als consument interpreteren van patronen in al verzamelde gegevens. Zodoende speelt statistiek een essentiële rol speelt in de wetenschapscyclus. Zelfs in situaties die relatief eenvoudig lijken kent statistiek de nodige valkuilen. Dit wordt in deze inleiding geïllustreerd aan de hand van een voorbeeld rond verkeersongelukken op kruispunten.[13]

---

[1] Iclon, Universiteit Leiden, corresponding author, koppmgm@iclon.leidenuniv.nl
[2] Tilburg University, Utrecht University, [3] Vrije Universiteit Amsterdam, [4] Rijksuniversiteit Gronningen, [5] Hoge School Utrecht, [6] Freudenthal Instituut, Utrecht University, [7] SLO, Amersfoort, [9] Hoge School van Amsterdam, [10] Iclon, Universiteit Leiden, [12] Stedelijk Gymnasium Leiden

[13] Dit voorbeeld is gebaseerd op de presentatie van Casper Albers bij de Koninklijke Nederlandse Academie der Wetenschappen op 24 mei 2022, https://www.knaw.nl/nl/bijeenkomsten/wiskunde-en-statistiek-het-voortgezet-onderwijs.

Het onderzoeksproces begint standaard met een onderzoeksvraag, zoals bijvoorbeeld de vraag naar de relatie tussen het type kruisingen en het aantal verkeersongelukken. Om deze onderzoeksvraag te beantwoorden worden data uit een bepaalde populatie verzameld. We kunnen binnen de data nu labels, aantallen of hoeveelheden toekennen; je typeert een kruispunt (bijvoorbeeld of er wel of geen stoplichten aanwezig zijn), telt een bepaald aantal passerend verkeer per dag, et cetera. Dat toekennen leidt tot verschillende uitkomsten per kruispunt, want niet elk kruispunt kent evenveel verkeer en een rotonde is een ander type kruispunt dan een T-splitsing. Dat aantal passerend verkeer wordt om die reden een *variabele* genoemd – het aantal kan variëren.

De eerste vraag die trouwens altijd gesteld moet worden is welke data je nodig hebt om de onderzoeksvraag te kunnen beantwoorden, of andersom, welke vragen met de beschikbare data mogelijke beantwoord zouden kunnen worden. Daarbij is de manier waarop de data verzameld wordt van belang.

Het is onmogelijk om alle kruispunten te meten. We observeren slechts een klein deel van de populatie, en dit geobserveerde deel noemen we de steekproef. Op basis van de steekproef kunnen we een dataset maken. Een dataset is een tabel met voor elke geobserveerd element in de steekproef de waarde voor elke variabele. De dataset ziet er dan bijvoorbeeld als volgt uit. Voor een steekproef van 120 kruispunten is gemeten wat voor type kruispunt het is, hoeveel ongevallen er waren in 2023 en hoeveel verkeer er was in datzelfde jaar.

Kruispunt	**Type**	**Verkeersongevallen in 2023**	**Verkeersdeelnemers (in miljoenen) in 2023**
1	Rotonde	18	3,23
2	Stoplichten	7	0,74
3	Stoplichten	35	2,45
…	…	…	…
120	Rotonde	5	1,94

Vervolgens wordt de dataset geanalyseerd om antwoorden te vinden op de onderzoeksvraag naar de relatie tussen het type kruising en het aantal verkeersongelukken. Dit gaat verder dan alleen statistische berekeningen, zoals het bepalen van het gemiddeld aantal ongelukken per type kruispunt. Ook als er in de dataset minder ongelukken zijn bij gelijkwaardige kruisingen met stoplichten dan bij rotondes, kan je niet direct de conclusie trekken dat rotondes onveiliger zijn. Immers, wat was bijvoorbeeld de reden om juist daar rotondes te maken? Wellicht is dat gebeurd omdat rotondes beter geschikt zijn voor kruispunten met veel verkeersintensiteit (en, daarmee samenhangend, meer mogelijkheden tot ongelukken).

Omdat een voorlopig antwoord altijd weer een vervolgvraag kan oproepen, spreekt men ook wel van de empirische onderzoekscyclus. Kennis van zo'n onderzoeksproces en de empirische cyclus is ook nodig als je statistische resultaten van anderen wilt interpreteren en beoordelen. De beschrijvingsvragen en verklaringsvragen bij aanvang van het onderzoeksproces komen niet uit de lucht vallen, maar komen voort uit context, waarneming en kennis.

Sectie 2 gaat dieper in op de empirische onderzoekscyclus en onderzoek doen in een bepaalde context, en beschrijft de rol van de statistiek daarin. In de daaropvolgende secties geven we een overzicht van de belangrijkste onderwerpen in de statistiek die ons inziens aandacht verdienen in de bovenbouw van het voortgezet onderwijs en/of die nodig zijn voor het schetsen van het landschap van de statistiek.

Sectie 3 geeft weer hoe we in de statistiek de data, bijvoorbeeld het aantal verkeersongelukken per kruispunt, het beste kunnen beschrijven. Zo'n beschrijving kan op verschillende manieren, zoals een figuur of tabel, of door middel van enkele getallen die (hopelijk) de essentie van de data uitdrukken, zoals een gemiddelde en spreidingsmaat van de waarden. Deze vorm van statistiek heet *beschrijvende statistiek*.

In Sectie 4 kijken we naar de beschrijvende statistiek van twee of meer variabelen. Meestal worden in een dataset meerdere variabelen tegelijk gemeten, variabelen waarvan men vermoedt dat deze met elkaar samenhangen of invloed hebben op deze samenhang en het begrijpen ervan. De bekendste voorbeelden van beschrijvende statistiek van twee variabelen zijn het verschil tussen gemiddelden van twee groepen en samenhang tussen twee variabelen. In ons voorbeeld naar de relatie tussen het type kruising en het aantal verkeersongevallen zijn dat bijvoorbeeld het (gemiddelde) verschil in aantal verkeersongelukken tussen kruisingen met stoplichten en kruisingen zonder stoplichten, en de correlatie (een maat voor samenhang) tussen het aantal verkeersongelukken op een kruising en het aantal verkeersgebruikers dat per uur de kruising passeert. Samenhang kan op verschillende manieren beschreven worden, bijvoorbeeld in een figuur of correlatiecoëfficiënt. Als je een samenhang gevonden hebt tussen variabelen, wil je daar doorgaans ook een causale verklaring aan hangen. Het is belangrijk om te onthouden dat correlatie en causaliteit twee verschillende concepten zijn en het is doorgaans niet mogelijk om op basis van enkel een dataset causale uitspraken te doen.

Vervolgens gaan we in Sectie 5 in op het concept *toeval*. Had de steekproef in ons voorbeeld uit 120 andere kruispunten bestaan, dan had de dataset er anders uitgezien. Omdat we uiteindelijk niet specifiek geïnteresseerd zijn in onze steekproef, maar in de gehele populatie waaruit deze getrokken is, is het belangrijk dat we kunnen inschatten hoe groot de rol van het toeval is.

Tenslotte behandelen we in Sectie 6 de *inferentiële statistiek*. Hierbinnen trek je op basis van de steekproef een conclusie over de gehele populatie, bijvoorbeeld dat kruispunten met rotondes veiliger zijn dan met stoplichten. Van de onderwerpen die we behandelen, is inferentiële statistiek het meest ingewikkeld, zowel vanuit wiskundig oogpunt als vanwege de interpretatie van de statistische analyses.

Samenvattend: in dit paper schetsen we een overzicht van de rol die statistiek speelt in de empirische onderzoekscyclus. We noemen daarbij de onderwerpen waarvan wij vinden dat ze aan bod kunnen komen in het onderwijs van de bovenbouw van havo en vwo, maar kijken ook iets verder in het statistische landschap. In Sectie 7 geven we enkele adviezen.

2 De empirische onderzoekscyclus

De empirische onderzoekscyclus of PTO-cyclus kan versimpeld worden samengevat in de drie fasen Probleem, Theorie, en Observatie/Onderzoek om het probleem aan te pakken. Eigenlijk heeft de cyclus, net als een cirkel, geen begin. Het probleem of daarmee samenhangend de beschrijvings- of verklaringsvraag komt voort uit observaties of vorig onderzoek, en ontstaat dus niet in isolement. Zo kan de vraag naar de relatie tussen het type kruisingen en het aantal verkeersongelukken bijvoorbeeld ontstaan zijn door een observatie of vermoeden dat rotondes veiliger zijn, en/of door theorie gedreven onderzoek dat laat zien dat de kans op ongelukken afneemt als de snelheid van voertuigen afneemt.

Als het probleem of de onderzoeksvraag helder is, verkregen door observatie of afgeleid uit theorie, dan worden data verzameld om tot een antwoord te komen. Data kunnen nooit los worden gezien van de context waarin ze tot stand zijn gekomen. De context kan hier worden gezien als de wereld om ons heen in al haar facetten. Het is aan de wetenschapper om data te destilleren uit de context, met de onderzoeksvraag in het achterhoofd. Toegespitst op ons voorbeeld, welke

wegen en weggebruikers zijn van belang, wat voor kruisingen kunnen we onderscheiden, wat zien we als verkeersongelukken, et cetera. Er is niet één of 'beste' manier; onderzoekers kunnen verschillen in hun formulering van de vraag, de theorie die ze aanhangen, als ook de onderzoeksmethode die ze gebruiken om hun onderzoeksvraag te beantwoorden.

Waar de PTO-cyclus als de traditionele empirische onderzoekscyclus kan worden beschouwd, zien we tegenwoordig steeds meer toepassingen van *data science* en *big data*. Het gaat hierbij om het analyseren van gegevens uit verschillende al bestaande bronnen, waarvan sommige mogelijk niet zorgvuldig zijn verzameld, of zijn verzameld voor een ander doel dan de huidige toepassing. Vaak gaat het hierbij om zeer grote hoeveelheden gegevens. Doordat de dataverzameling meestal niet het gevolg is van theorie is de data-analyse vaak *exploratief*. Er zijn vaak veel data voorhanden waarvan nog moet worden vastgesteld of ze van nut zijn bij het beantwoorden van een onderzoeksvraag. Dat maakt het leren analyseren, interpreteren en conclusies trekken uit big data en exploratieve data-analyse—inclusief leren wat er niet uit mag worden geconcludeerd—belangrijke vaardigheden. Bij ondoordacht gebruik van data kan het anders zomaar gebeuren dat een gevonden samenhang toevallig was en/of geen praktische betekenis heeft. Waar het bijvoorbeeld duidelijk is dat de sterke samenhang over de tijd tussen de populariteit van de voornaam 'Sunny' in de VS en de hoeveelheid door de zon opgewekte energie in Egypte betekenisloos is (zie https://www.tylervigen.com/spurious-correlations voor mooie voorbeelden van betekenisloze samenhangen zoals deze), is dat lang niet altijd duidelijk. Om te beoordelen of een gevonden samenhang betekenisvol is is naast statistiekkennis dan ook inhoudelijke kennis van de context nodig.

Naast het geheel zelf doorlopen van de PTO-cyclus (d.w.z., je doet zelf onderzoek waarbij je data verzamelt om je eigen onderzoeksvraag te beantwoorden) en het analyseren van bestaande data, kennen we ook het evalueren van onderzoeksuitkomsten waarbij statistiek is gebruikt. Mensen komen veelvuldig in aanraking met zulke onderzoeksuitkomsten, in kranten of social media. En omdat deze onderzoeksuitkomsten regelmatig gekleurd worden gepresenteerd, is het zelfstandig kunnen evalueren ervan cruciaal.

In veel toepassingen worden data verkregen uit onderzoek. Onderzoek doen betekent dat je op een systematische manier een bepaalde onderzoeksvraag bestudeert. Er zijn veel manieren om onderzoek te doen. Je kan de (wetenschappelijke) literatuur induiken, observaties doen, interviews afnemen, of een experiment doen. Voordat je je onderzoek uitvoert, is het belangrijk dat je goed nadenkt over hoe je je onderzoek opzet en hoe je je data verzamelt. Onderzoek is als een kaartenhuis; als de onderzoeksvraag, onderzoeksopzet en methode van dataverzameling niet goed zijn, dan kun je met statistiek geen betekenisvolle conclusies meer trekken na data-analyse. Elke methode van onderzoek heeft zijn eigen kenmerken en valkuilen, en gaat gepaard met grotendeels eigen methoden van data-analyse. Voor hier voert het te ver om in te gaan op deze onderzoeksmethoden, maar we benadrukken graag dat het onderzoeksproces één geheel is; statistiek en data-analyse kan niet los worden gezien van context, vraagstelling, en het onderzoek waarmee de data zijn verkregen.

Om de PTO-cyclus te kunnen doorlopen, om data te kunnen analyseren, of om artikelen of nieuwsberichten te kunnen begrijpen waar statistiek is gebruikt, is natuurlijk ook kennis nodig van de belangrijkste onderwerpen in de statistiek. Deze onderwerpen bespreken we in secties 3 t/m 6.

3 Beschrijvende statistiek: één variabele

We beginnen met relatief de meest eenvoudige vorm van statistiek, de beschrijvende statistiek van één variabele.

Aan een eigenschap van een object of wezen kan vaak een getal worden toegekend, een proces dat *meten* wordt genoemd. We spreken over variabelen, zoals bijvoorbeeld de lengte in meter van giraffen in de dierentuin Beekse Bergen of het aantal verkeersongelukken per kruising in het jaar 2022. Als het aantal mogelijke waarden van een variabele stapsgewijs varieert, bijvoorbeeld een aantal verkeersongelukken, dan is er sprake van een *discrete variabele*. Kan een variabele iedere waarde op een interval aannemen, bijvoorbeeld de lengte van een giraf, dan spreken we over een *continue variabele*.

Als de variabele bij een aantal objecten (kruisingen) of wezens (giraffen) is gemeten kan er een *verdeling* van de variabele worden geconstrueerd. Dat kan zowel grafisch als getalsmatig, afhankelijk van wat je hebt gemeten. Er zijn veel verschillende grafische voorstellingen van verdelingen, maar de bekendste daarvan is de *histogram* en varianten daarvan. Op de *x*-as vind je dan de waarde, en op de *y*-as het aantal keer dat deze waarde voorkomt, ook wel de (absolute of relatieve) frequentie genoemd. In geval van relatieve frequenties hebben we te maken met een bijzondere verdeling waarvan de oppervlakte gelijk is aan 1.

Bij een grafische voorstelling van een verdeling worden met name drie aspecten van belang geacht; *centrum*, *spreiding* en *vorm*. Centrum verwijst naar de meest kenmerkende score van de verdeling, zoals (i) top, ook wel modus genoemd, (ii) halverwege ofwel mediaan, (iii) zwaartepunt ofwel gemiddelde. In de statistiek wordt het gemiddelde het vaakst gebruikt als centrummaat van een verdeling. De spreiding van een verdeling heeft betrekking op hoeveel de waarden in de verdeling uit elkaar liggen. Als alle objecten of wezens dezelfde waarde hebben op de variabele dan is er geen spreiding.

De vorm heeft betrekking op hoe de verdeling er uitziet. In de praktijk zien we vaak ééntoppige verdelingen (d.w.z., met één modus), die soms nagenoeg symmetrisch zijn (dan zijn modus, mediaan en gemiddelde ongeveer gelijk), rechtsscheef met een staart naar rechts (in de regel geldt dan dat modus < mediaan < gemiddelde) of linksscheef met een staart naar links (in de regel gemiddelde < mediaan < modus). Enkele ééntoppige verdelingen komen vaak voor in de statistiek, of omdat ze bij benadering worden geobserveerd in de natuur of omdat ze heel nuttig blijken te zijn bij begrijpen en onderzoeken van verschijnselen. Voorbeelden daarvan zijn de symmetrische en klokvormige normale verdeling, en de discrete binomiale verdeling.

De getalsmatige voorstelling is een lijst met voorkomende waarden en hun (relatieve) frequenties, of een functievoorschrift als het gaat om een bekende verdeling zoals bijvoorbeeld de eerdergenoemde normale verdeling of binomiale verdeling. Elke waarde in de verdeling heeft ook een percentielscore, gelijk aan het percentage van waarden in de verdeling die even groot of lager is. Centrum- en spreidingsmaten kunnen ook worden berekend. Naast het gemiddelde, mediaan (50^e percentielscore) en modus als centrummaat kennen we als belangrijkste spreidingsmaten de variantie, standaardafwijking, bereik, en interkwartielafstand (IQR). De variantie van een verdeling is de gemiddelde gekwadrateerde afstand tot het gemiddelde van die verdeling, en de standaardafwijking is daar de wortel van. Het bereik of spreidingsbreedte is de afstand tussen het maximum en het minimum van de verdeling en de IQR is de afstand tussen het 25^e en 75^e percentielscore van de verdeling.

Een verdeling kan ook één of meer zogenoemde *uitbijters* bevatten. Een uitbijter is een waarde die of veel lager of veel hoger is dan bijna alle andere waarden in de verdeling. Als de lengte van bijna alle giraffen in de Beekse Bergen tussen de 4 meter en 5,25 meter ligt, dan is een giraffe met lengte 6,25 meter een uitbijter. Uitbijters hebben een grote invloed op de getalsmatige voorstelling van een verdeling. In het voorbeeld zijn het gemiddelde en standaardafwijking van de lengte van giraffen in de Beekse Bergen aanmerkelijk hoger met dan zonder deze uitbijter. Vaak is er een bijzondere reden voor een uitbijter (bijvoorbeeld, de giraffe van 6,25 meter is van een

andere soort; de Beekse Bergen kent drie soorten giraffen), en op basis van de bijzondere reden kan worden besloten een uitbijter wel of niet te verwijderen uit de verdeling. Dat mag echter nooit zomaar, en al helemaal niet omdat het de onderzoeker beter uit zou komen.

4 Beschrijvende statistiek: twee of meer variabelen

Een veelvoorkomend voorbeeld van beschrijvende statistiek van twee variabelen is het vergelijken van twee verdelingen. Meestal is men dan geïnteresseerd in het verschil in gemiddelden van deze verdelingen. De ene variabele is dan continu, de andere variabele heeft twee mogelijke waarden (binair). Toegespitst op ons voorbeeld kunnen we het gemiddeld aantal ongelukken (de continue variabele) vergelijken van gelijkwaardige kruisingen en rotondes (type kruising, de binaire variabele). Een getalsmatige beschrijving die samenvat hoe sterk het effect is van de ene variabele op een andere variabele is bijvoorbeeld de effectgrootte Cohen's d.

Een ander veelvoorkomend voorbeeld is *samenhang* tussen twee variabelen, zoals bijvoorbeeld de samenhang tussen het aantal verkeersongelukken op een kruising en het aantal verkeersgebruikers dat per uur de kruising passeert. Er kan onderscheid gemaakt worden tussen de richting van de samenhang (positief of negatief) en de sterkte van de samenhang. De beschrijving kan woordelijk, visueel of getalsmatig gedaan worden. Een woordelijke beschrijving van de samenhang heeft vaak de vorm van een uitspraak als: "als X groot is, dan is Y over het algemeen klein". Dit is een voorbeeld van een negatieve samenhang. Sterkere samenhang kan bijvoorbeeld uitgedrukt als: "als X groot is, dan is Y bijna altijd klein". Een figuur als de puntenwolk (scatterplot) kan de samenhang ook weergeven. De bekendste getalsmatige beschrijving van de samenhang van twee variabelen is de correlatiecoëfficiënt, met waarden tussen -1 (perfecte negatieve lineaire samenhang) en 1 (perfecte positieve lineaire samenhang), met de waarde 0 die de afwezigheid van lineaire samenhang aangeeft. De correlatiecoëfficiënt is er in vele vormen, afhankelijk van wat en hoe je precies de variabelen meet (bijvoorbeeld phi, Spearman's ρ en Pearson's coëfficient).

Ook als het gaat om het beschrijven van samenhang dient rekening te worden gehouden met uitbijters, want bijvoorbeeld de waarde van de Pearson's correlatiecoëfficiënt is daar zeer gevoelig voor. Eén zeer grote onveilige kruising met relatief veel verkeersgebruikers en tegelijkertijd ook veel verkeersongelukken kan ervoor zorgen dat de correlatie sterk positief is, ook als er in de dataset zonder deze kruising geen samenhang is tussen deze variabelen.

Bij het bepalen van samenhang tussen variabelen willen onderzoekers vaak rekening houden met andere variabelen die deze samenhang kunnen beïnvloeden. Toegespitst op ons voorbeeld, bij de vergelijking tussen het gemiddeld aantal ongelukken bij gelijkwaardige kruisingen en rotondes zullen we ook rekening willen houden met aantal verkeersgebruikers dat per uur gebruik maakt van de kruising. Immers, als er een verschillend aantal verkeersgebruikers zijn op rotondes dan op gelijkwaardige kruisingen, dan kan een verschil in gemiddelde het gevolg zijn van het aantal verkeersgebruikers in plaats van de veiligheid van het type kruising. Juist het rekening houden met andere variabelen die de samenhang kunnen beïnvloeden is van belang bij het evalueren van krantenartikelen of nieuwsberichten waarbij statistiek is gebruikt.

Ook komt het vaak voor dat een onderzoeker één variabele wil verklaren of voorspellen met een aantal andere variabelen (ook predictoren genoemd), bijvoorbeeld het aantal verkeersongelukken met type kruising, het aantal verkeersgebruikers, aanwezigheid stoplichten, locatie kruising, et cetera. In dat geval wordt vaak een *regressie-analyse* gebruikt. Bij een variabele die aan de hand van één predictor voorspeld wordt, wordt dan een lineair model gemaakt, de regressielijn. Zijn er meerdere predictoren dan spreken we van multiple of meervoudige regressie. Regressie-analyse vomt ook de basis voor machine learning waarbij computers getraind worden om

op basis van gegeven input-output paren een model te construeren dat gebruikt kan worden om de uitkomst bij nieuwe input te voorspellen en/of om verbanden tussen grootheden te vinden.

Het beschrijven van de samenhang is overigens meestal niet voldoende. Het is ook wenselijk de samenhang te begrijpen. Samenhang hoeft geen causaal verband te impliceren, en kan ook wijzen op een gemeenschappelijke oorzaak. Bijvoorbeeld, de hoeveelheid chocola die per hoofd van de bevolking wordt geconsumeerd correleert uitstekend met het aantal Nobelprijzen per hoofd van de bevolking. Deze correlatie is uiteraard niet causaal, maar wordt veroorzaakt door een gemeenschappelijke factor: welvaart. Het kunnen vaststellen van causaliteit hangt samen met de gebruikte onderzoeksopzet en is zelden eenvoudig. Bij machine learning is deze waarschuwing extra relevant.

5 Toeval en kans

Statistiek is onlosmakelijk verbonden met *toeval* en *kans*. Voldoende begrip van toeval en kans is noodzakelijk om statistische gegevens goed te kunnen interpreteren. In de waarschijnlijkheidsrekening en statistiek zeggen we dat een gebeurtenis een toevallige uitkomst heeft als de uitkomst van de gebeurtenis niet van tevoren vaststaat. Bijvoorbeeld aan de uitkomst van het gooien met één muntstuk. Soms is de uitkomst onzeker, maar weet je wel de kans op een bepaalde uitkomst, zoals je weet dat de kans op kop en ook op munt bij een eerlijke munt gelijk is aan 1/2. Toeval en kans spelen echter ook een belangrijke rol bij onderzoeken in de empirische cyclus. Om weer terug te komen op ons voorbeeld naar de relatie tussen het type kruisingen en het aantal verkeersongelukken; het is belangrijk te beseffen dat kruisingen verschillen in het aantal verkeersongelukken. Misschien is de keuze voor juist die kruisingen ook door toeval tot stand gekomen. En met dit toeval dient rekening te worden gehouden bij uitspraken over verkeersongelukken in de hele populatie van kruisingen.

Alle mogelijke uitkomsten van een gebeurtenis kunnen worden weergegeven in een kansverdeling, vergelijkbaar met de frequentieverdeling die we zagen bij de beschrijvende statistiek. De som van alle kansen is één, en kansen op uitkomsten variëren van nul tot één. Een bekende kansverdeling is de binomiale verdeling. De binomiale verdeling wordt gebruikt als er sprake is van een gebeurtenis met slechts twee mogelijke uitkomsten, waarbij één van de uitkomsten vaak ‘succes’ wordt genoemd, en waarbij dit experiment een aantal keren herhaald wordt. Bijvoorbeeld: de kansen van het aantal keren ‘zes’ bij het 15 keer gooien met een dobbelsteen. Kenmerkend voor een binomiale verdeling is dat de kans op succes constant is én dat de uitkomsten van verschillende worpen *onafhankelijk* van elkaar zijn. Omdat onafhankelijkheid een centrale aanname is van veel toepassingen van de statistiek, is het begrip van (on)voorwaardelijke kans en onafhankelijkheid cruciaal voor het begrip van de statistiek.

De bekendste kansverdeling is zonder meer de normale verdeling. Waar de binomiale verdeling een discrete verdeling is waarbij alleen de gehele getallen van 0 tot N (successen) voor kunnen komen bij N herhaalde onafhankelijke gebeurtenissen, is de normale verdeling continu. De normale verdeling heeft de vorm van een klok, symmetrisch rond het gemiddelde (die in dit geval gelijk is aan mediaan en modus). Het belang van de normale verdeling ligt in de belangrijkste stelling van de waarschijnlijkheidsrekening, de *centrale limietstelling*. Bij een groot aantal onafhankelijke trekkingen uit een onbekende verdeling wordt steeds het gemiddelde bepaald; deze gemiddelden zullen niet steeds gelijk zijn. De stelling houdt in dat deze gemiddelden goed passen in een normale verdeling, ongeacht de vorm van de onbekende verdeling. Omdat veel statistische toepassingen uitgaan van veel onafhankelijke trekkingen uit een populatieverdeling (vaak met onbekende vorm), komt deze normale verdeling goed van pas in deze toepassingen. Als eenvoudig

voorbeeld, de verdeling van het gemiddeld aantal zessen na 100 worpen wordt verrassend goed benaderd door een klokvormige normale verdeling.

Een andere belangrijke wet uit de kansrekening die noodzakelijk is voor goed begrip van de statistiek is de zogenaamde *wortel-n-wet*. Deze wet zegt dat als het aantal onafhankelijke waarnemingen op basis waarvan je een gemiddelde berekent *n* keer zo groot wordt, de standaardafwijking van dat gemiddelde $1/\sqrt{n}$ keer zo groot is als de standaardafwijking van de oorspronkelijke waarnemingen. Direct gevolg van de stelling en de wet is dat bij een groot aantal steekproeven met steekproefomvang *n* uit een populatie, de steekproevengemiddelden bij benadering een normale verdeling hebben (centrale limietstelling) met standaardafwijking gelijk aan de standaardafwijking van de populatie gedeeld door $\sqrt{n}$ (wortel-n-wet).

6 Inferentiële statistiek

Met de inferentiële statistiek trek je op basis van de steekproef een conclusie over de gehele populatie. De inferentiële statistiek is een toepassing van de wetmatigheden die in de vorige sectie over toeval en kans zijn beschreven. Als via een steekproef een uitspraak gedaan moet worden over een populatie, dan is dat het meest voor de hand liggend als het om een *toevallige (aselect) getrokken steekproef* (simple random sample) gaat. Dit betekent dat Erik en Erika een even grote kans hebben om in de steekproef voor te komen én dat de trekking *onafhankelijk* is, d.w.z. dat de gebeurtenis dat Erika in de steekproef zit niet afhangt van het wel of niet voorkomen van Erik in de steekproef.

Als gevolg van een enkelvoudig toevallig (aselect) getrokken steekproef uit de populatie kan de verdeling van een statistische grootheid worden afgeleid, bijvoorbeeld de verdeling van het steekproefgemiddelde, maar ook de verdeling van andere grootheden zoals bijvoorbeeld een correlatiecoëfficiënt. Toegepast op de verdeling van het steekproefgemiddelde, de zogenaamde steekproevenverdeling van het gemiddelde, volgt uit de eerder besproken centrale limietstelling en de wortel-n-wet dat deze goed benaderd wordt door de normale verdeling (bij voldoende grote steekproefomvang), met gemiddelde gelijk aan het populatiegemiddelde en standaardafwijking gelijk aan de standaardafwijking van de populatie gedeeld door door $\sqrt{n}$.

Bij *betrouwbaarheidsintervallen* ga je uit van een gevonden steekproefgemiddelde en bereken je op basis daarvan een interval. Vaak wordt gewerkt met het 95%-betrouwbaarheidsinterval, dat wil zeggen dat als je de procedure heel vaak zou herhalen, in 95% van de gevallen het aldus berekende interval het populatiegemiddelde zal omvatten. Dit interval wordt smaller, volgens de wortel-n-wet, naarmate de steekproef groter wordt.

Een in de wetenschap veelgebruikte vorm van inferentie is het *toetsen van een hypothese*. Dat begint bij een aanname of *nulhypothese* over bijvoorbeeld het verschil van twee populatiegemiddelden, en rekent dan de kans uit op de waarde van het gevonden verschil tussen de steekproefgemiddelden of extremer, gegeven dat de nulhypothese waar is. Deze kans wordt *p-waarde* genoemd. Als de *p*-waarde klein is, zeg kleiner dan 0,05, dan wordt de nulhypothese verworpen, en in ons voorbeeld concludeer je dan dat de twee populatiegemiddelden van elkaar verschillen. Hoewel breed toegepast is deze procedure niet onproblematisch in de zin dat ze geen antwoord geeft op de onderzoeksvraag of de nulhypothese waar is: er wordt immers gerekend onder de aanname dat deze waar is.

De vraag hoe waarschijnlijk het is dat een hypothese waar is wordt geadresseerd in de Bayesiaanse benadering van de statistiek. In deze benadering beginnen we met twee hypotheses. Nadat we de data hebben verzameld wordt de kans uitgerekend op precies deze data, onder elk van de twee hypotheses. De verhouding van die kansen heet de likelihood ratio, en dit getal vertelt

ons welke van de twee hypotheses de data het beste verklaart, en met welke factor. Dit geeft niet direct de kans op de hypotheses: daarvoor is het nodig dat de onderzoeker aangeeft hoe groot hij of zij de kans op de hypotheses oorspronkelijk inschatte. Samen met die oorspronkelijke (prior) kans en de gevonden likelihood ratio kan dan de zogenaamde posteriori (achteraf) kans op de hypotheses worden berekend, die van persoon tot persoon kan verschillen.

7 Statistiekonderwijs in de bovenbouw van havo en vwo

Statistiek en kansrekening vormen samen een groot en in elkaar verweven domein met relevante kennis en toepassingen voor de leerlingen van havo en vwo. Bij het profielwerkstuk, bij andere vakken maar ook in veel vervolgstudies, op hbo en universiteit, zal kennis op dit domein goed van pas komen, en ook bij het analyseren van bestaande data en bij het beoordelen van statistisch werk van anderen. Daarnaast heeft het voortgezet onderwijs als doel leerlingen voor te bereiden op het functioneren in de maatschappij, die steeds meer data-gestuurd is. Leerlingen moeten leren statistische informatie in media en het dagelijks leven te herkennen, te lezen, te interpreteren, kritisch te beoordelen en erover te communiceren: ze moeten kortgezegd statistisch geletterd worden. Naast gezond verstand is daar ook statistisch gereedschap voor nodig. De hamvraag is natuurlijk hoe we de leerlingen daar het best op kunnen voorbereiden door middel van ons statistiekonderwijs.

Er zijn hierbij twee duidelijke rode vlaggen. Ten eerste, statistiek is geen kwestie van op de juiste knoppen van een softwarepakket drukken en dan automatisch regels toepassen als “de *p*-waarde is kleiner dan 0,05 dus de nulhypothese is onwaar”. Ten tweede, statistiek is ook niet het veelvuldig uitrekenen van ‘sommetjes’. Er moet begrip zijn over wat er gebeurt, en er moet een link met de wereld om ons heen zijn. We raden hierbij aan om in het onderwijs te beginnen met de beschrijvende statistiek van één variabele (verdeling, gemiddelde, percentielen en spreiding), in de context van aansprekende voorbeelden die bij voorkeur zijn gekoppeld aan de inhoud van andere vakken en de leefomgeving van de leerling. Daarbij worden ook krantartikelen en berichten uit social media gebruikt waarin conclusies getrokken worden op basis van statistiek, waarbij niet gerekend hoeft te worden maar de conclusies kritisch moeten worden geëvalueerd. Deze benadering kan ook worden gevolgd bij het achtereenvolgens bespreken van samenhang tussen twee variabelen, toeval en kans, en inferentiële statistiek.

Omdat het begrijpen van inferentiële statistiek (ook voor andere wetenschappers) lastig is, is het raadzaam om gebruik te maken van hulpmiddelen. We denken hierbij aan *simulaties* om de variatie die kan optreden bij het trekken van steekproeven zichtbaar te maken via een steekproevenverdeling. Ook het opstellen van een betrouwbaarheidsinterval kan met simulaties zichtbaar en begrijpelijk gemaakt worden via software die relatief eenvoudig te bedienen is, zoals VUStat, StatKey, Codap, et cetera. In combinatie met kansrekening kunnen deze hulpmiddelen ingezet worden om begrip te kweken van de centrale limietstelling, de wortel-n-wet en de inferentiële statistiek. Het is daarbij belangrijk dat leerlingen de systematische aanpak bij onderzoek doen leren kennen en herkennen. Het is raadzaam om de empirische onderzoekscyclus te illustreren aan de hand van voorbeelden uit bestaande vakken of al voltooide profielwerkstukken, ter voorbereiding op hun eigen profielwerkstuk en het hogere onderwijs.

Machine learning leent zich ook uitstekend voor behandeling op school: er zijn tamelijk gebruiksvriendelijke softwarepakketten, en machine learning verbindt de gevestigde theorie met de moderne ontwikkelingen. Maar ook hier geldt: het mag geen kwestie van alleen maar op knoppen drukken worden, en het is belangrijk dat leerlingen weten wat ze precies aan het doen zijn.

Kort samengevat zijn zijn onze suggesties voor het voortgezet onderwijs: besteed aandacht aan alle fasen van de empirische onderzoekscyclus, leer kernvragen om eigen onderzoek en de statistische informatie in onder andere media en dagelijks leven kritisch te bevragen, en gebruik daarbij, naast common sense, ook een beperkt aantal kernconcepten en statistische tools, en wees ten slotte niet bang om aan te sluiten bij recente ontwikkelingen.