**Toeval in de greep door simulatie.**

Douwe Wielenga en Piet van Blokland

Omdat tegenwoordig statistiek bij VWO A/C alleen nog in het schoolexamen zit, is een moderne didactische aanpak en toetsing mogelijk waarbij simulatie een dragende rol speelt.

Sinds juni 2021 zijn de VUStat apps met bijbehorend lesmateriaal "Toeval in de greep" door de NVvW website beschikbaar gemaakt. Het adres met docenten-informatie is: **toevalindegreep.nl/doc**.

Daar staan ook de doorverwijzingen naar leerlingenmateriaal voor VWO 4 en 5. Het lesmateriaal is het resultaat van 5 jaar ontwikkelen op scholen en is vrij te gebruiken en eventueel aan te passen.

Het is goed mogelijk alleen een deel te gebruiken. Dit alles is de aanleiding voor dit artikel in Euclides.

Piet van Blokland is de auteur van VUGrafiek en VUStat. Douwe Wielenga is de auteur van het lesmateriaal. Beiden startten in 1971 de wiskunde afdeling van de nieuwe lerarenopleiding VL-VU.

**Inleiding.**

Statistiek onderwijs in het VWO begint meestal met een fase in de vierde klas waarin op allerlei manieren informatie uit complete gegevensbestanden zichtbaar wordt gemaakt. Daarmee kun je dan vragen over de hele bekeken populatie al of niet beantwoorden. Daar gaat dit artikel niet over.

Vervolgens stappen we over op de fase waarin het gegevensbestand een **steekproef** is uit een grotere populatie. Op grond daarvan willen we uitspraken doen over de populatie zelf. Maar dan komt er onzekerheid bij, veroorzaakt door het toeval dat er bij het trekken van die (aselecte) steekproef is ingeslopen. Dus is het van belang deze fase te beginnen met een onderzoek van de eigenschappen van het toeval. Dat kan tegenwoordig heel goed met simulaties van toevalsprocessen.

**Simulatie en inzicht in de werking van het toeval.**

Simulaties geven inzicht in het basis idee: "grotere steekproeven dringen met hun gemiddelde het toeval terug". Je krijgt het toeval meer in de greep. Dat is de wet van de grote aantallen. Het kansbegrip wordt daaraan opgehangen.

Voor dat basis idee zit in het pakket apps van VUStat een voorbeeld van een toevalsproces waarbij het niet evident is wat de kans op succes is: het is het aloude TV-spel met de drie deuren en een prijs achter één ervan. Je mag een deur kiezen en de spelleider, die weet waar de prijs zit, opent nu van de andere twee deuren net die deur waar geen prijs achter zit.

Wat doe je: wissel je nog van deur of blijf je bij je oorspronkelijke keuze?

Leerlingen kunnen bij herhaling experimenteren met die keuze, en hun resultaten worden via internet in een gezamenlijke grafiek omgezet. Daaraan zie je al hoe het succespercentage bij toenemende steekproefomvang naar een stabiele waarde lijkt te gaan.



En dan laat je het simulatieproces helemaal door de computer overnemen en kun je experimenteren met de variabiliteit van het eindpercentage bij verschillende steekproefomvang. In eerste instantie kijk je naar de boxplot, en in tweede instantie naar de standaardafwijking. Zo wordt de wortel-n-wet experimenteel ontdekt: 25 keer zoveel werk levert 5 zo kleine variabiliteit.

Nu is men gewend aan het idee dat je met simulatie kansen kan verkrijgen.

Als je dit statistische verschijnsel niet goed in je denken hebt verankerd, kun je makkelijk met je intuïtie verkeerde conclusies gaan trekken. Het blijkt dat die verankering niet zo goed gaat bij veel leerlingen als je alleen woorden en formules gebruikt.

*Een voorbeeld van een verkeerde conclusie*

*(naar aanleiding van een artikel van Van Brederode & Meeter).*

Het centraal examen voor een vak (CE) is een steekproef met omvang 1. Maar het schoolexamen (SE) is een steekproef met een groter aantal cijfers waar het gemiddelde van genomen wordt.

We weten, uit die simulaties van zojuist, dat de variabiliteit van de verdeling van de CE cijfers in het land dus waarschijnlijk groter is dan die van de verdeling van de SE cijfers. Dat is ook zo: landelijk was vroeger de situatie zoals in dit plaatje. Bovendien lag het SE gemiddeld 0,2 punt hoger dan het CE,



Stel dat Douwe een vakniveau heeft dat 1 standaard-afwijking onder het landelijk gemiddelde ligt, consequent zowel bij het SE als bij het CE. De correlatie tussen SE en CE is in dat geval perfect.

Dan zie je aan het plaatje dat de verwachte score van Douwe bij het SE voor dat vak hoger zal liggen dan bij het CE. En bij Piet, die een niveau heeft dat 2 standaardafwijkingen boven het gemiddelde ligt, zal het verwachte SE juist lager liggen.

Een school met veel Douwes zal daarom naar verwachting gemiddeld een lager CE hebben dan SE, en een school met veel Pieten juist gemiddeld een hoger CE dan SE, blijkt uit een simpel rekensommetje.

Jarenlang is dit **statistische** verschijnsel **oorzakelijk** geïnterpreteerd: als je je concentreert op scholen met laag gemiddeld CE dan hebben die meestal een hoger SE dan CE; dus die zullen hun zwakke leerlingen wel een handje geholpen hebben op het schoolexamen! Daar moest de inspectie op controleren.

Zo zie je een ongewenst maatschappelijk gevolg van een gebrekkig statistisch inzicht.

Je kunt dit ook door simuleren laten zien als je de redenering niet helemaal vertrouwt, en je kunt daarbij verschillende waarden van de correlatie tussen SE en CE cijfers proberen. Dan blijkt: hoe zwakker de correlatie hoe sterker het effect. Mogelijk moeten de waargenomen data nog verder geanalyseerd worden om een definitief oordeel te kunnen vellen.

In het materiaal van de VWO lespakket Toeval in de greep komen dergelijke dingen bewust aan de orde en ze worden verder onderzocht, meestal met simulaties van toevalsprocessen. Er zijn wel meer voorbeelden waarbij de mens door zijn intuïtieve systeem-1-denken (Kahneman) eerst aan een oorzakelijk verband denkt, en daarna pas door zijn systeem-2-denken wordt gecorrigeerd, namelijk: het kan ook een statistische effect zijn, waar geen oorzaak voor nodig is. De wet-van-de-kleine-aantallen en regressie-naar-het-midden zijn daar voorbeelden van. Leerlingen moeten daarom enkele hoofdstukken lezen en samenvatten uit Kahnemans boek "Thinking fast and slow".

**Simulatie en inzicht in Verdeling, Schatten en Beslissen.**

Aan de basis van het didactisch concept ligt ook het volgende idee. Door computers veel aselecte steekproeven uit een populatie te laten trekken (simuleren) krijg je inzicht in de kansverdeling van het gemiddelde van zo'n steekproef. Hieronder staat een voorbeeld met een eerlijk muntstuk dat 100 keer gegooid wordt, en de steekproefgrootheid is het percentage successen.



Dat inzicht, dat begrip, ontstaat bij leerlingen maar moeizaam als je alleen woorden of formules gebruikt. Veel leerlingen blijven dan, gehandicapt, onbegrepen dingen met formules doen.

Als je kunt experimenteren met een computer die, -door jezelf gestuurd-, het resultaat van zeer veel herhaalde steekproeven toont in de vorm van een verdeling, dan zie je het toeval aan het werk en dan heeft een leerling een prothese die hem over die handicap heen helpt. Een leerling kan proefondervindelijk statistische wetmatigheden bij toevalsprocessen onderzoeken, cq. aantonen.

En nu kun je heel makkelijk overstappen op Schatten en Toetsen in de situatie waarin je slechts een (aselecte) steekproef hebt, en toch iets over de hele populatie wilt zeggen. Dat moet een leerling dan blijven doen met simulaties, en niet met formules. Daardoor wordt die voornoemde prothese steeds opnieuw gebruikt, als een soort hamer die het begrip er telkens dieper inslaat.

Om dezelfde reden is het verstandig de didactische opbouw concentrisch te maken: begrippen komen steeds terug en worden verder ingekleurd of uitgebouwd.

*Concentrisch*

Zo kun je bijvoorbeeld beginnen met het behandelen van het idee van hypothese-toetsen en daarna het betrouwbaarheidsinterval definiëren als de verzameling van alle hypotheses over de populatieparameter die nog acceptabel zijn in het licht van de waargenomen steekproef en de significantiedrempel van de gevolgde methode. Je hebt daarbij overigens wel een veronderstelling gedaan over het type kansmodel voor de populatie.

Dan merk je dat het wel heel veel werk is om de grenzen van dat betrouwbaarheidsinterval experimenterenderwijs te bepalen, en je valt terug op een snelle benaderingsmanier, "die wel ongeveer zal kloppen".

Maar in een latere fase kom je er op terug. De moderne aanpak van statistische problemen gebruikt **resampling** methodes. Nu is een veronderstelling over het type model voor de populatie nauwelijks meer nodig. De **bootstrap** simulatie "voegt bij herhaling wat toeval toe" aan de waargenomen steekproef en produceert zo een verdeling van de steekproefgrootheid. Het bijbehorende 95%-betrouwbaarheidsinterval blijkt nu met ongeveer 95% zekerheid de populatieparameter te bevatten. Leerlingen kunnen ook experimenteren met een app om dit kenmerk van de bootstrap methode te controleren onder verschillende voorwaarden.



Dat betrouwbaarheidsinterval is de schatting van de parameter van de populatie op grond van de steekproef. Daarna komt pas de beslissing of je een nulhypothese moet verwerpen of niet: zit die hypothese in het betrouwbaarheidsinterval of niet? Leerlingen hebben daar dan geen conceptuele moeite meer mee.

En in een weer latere fase bespreek je de samenhang tussen de alternatieve hypothese, de significantiedrempel en het betrouwbaarheidsinterval. Een derde serie opgaven sluit deze fase af.

Leerlingen kunnen daarna ook nog simulaties gebruiken om te onderzoeken hoe groot een steekproef moet zijn om, bij een gegeven significantiedrempel (= kans op fout type 1), de kans op fout type 2 beneden een zekere waarde te krijgen. Je moet je hersens erbij houden, maar door die simulaties begrijpen ze heel goed waar ze mee bezig zijn.

**Voorbeeld van een relevant onderzoek in de le**s.

Het idee is afkomstig van Alan Rossman en het onderzoek gaat over het *psychologisch referentie-effect.*

In een eerdere fase heb je in een groepsactiviteit al aandacht besteed aan **bias** bij een niet-echt-aselecte steekproef. En met een voorbeeld over het oorspronkelijk verkeerd opgezette onderzoek naar de werking van het Salk vaccin tegen polio, in 1954, wordt goed duidelijk dat een **dubbel blind aselect gecontroleerd test- en controlegroep** opzet echt nodig is om met statistiek een eventueel oorzakelijk verband te kunnen aantonen.

Nu verdeel je de klas aselect in twee groepen. De ene groep krijgt de vraag:

 *We vragen ons af of Nelson Mandela, de eerste president van Zuid Afrika na de apartheid,*

 *jonger of ouder was dan 16 jaar toen hij overleed.*

 *Geef jouw schatting van zijn leeftijd toen hij overleed.*

De andere groep krijgt de vraag:

 *We vragen ons af of Nelson Mandela, de eerste president van Zuid Afrika na de apartheid,*

 *jonger of ouder was dan 160 jaar toen hij overleed.*

 *Geef jouw schatting van zijn leeftijd toen hij overleed.*

En we voeren deze twee steekproeven in in de app "Twee gemiddelden vergelijken" en kiezen voor de bootstrap-methode. Dit is het resultaat:



De bootstrap methode levert de volgende verdeling van het verschil van de twee gemiddelden:



En de klas constateert een effect van de eerste rare onzin-zin van tussen de 10 en 27 jaren op de gemiddelde schatting van Mandela's leeftijd. En dat deze uitspraak voor 95% zeker is.

Het is een oorzakelijk effect want de opzet was met aselecte toekenning gerealiseerd.

Natuurlijk moet nog wel gediscussieerd worden over de steekproef zelf: welke populatie, en was de klas wel aselect? In elk geval is duidelijk dat de nulhypothese (geen invloed) wordt verworpen in de methodiek met een eenzijdige significantie van 2,5%.

Een leerling heeft eens een eigen-bedacht onderzoek opgezet en uitgevoerd in een basisschool om te kijken of de mening van de juf invloed heeft op de keuze van een kind.

Er zijn nog veel meer onderwerpen die een rol moeten spelen in een goede statistiek cursus, en vaak kun je daar de simulaties gebruiken voor de begripsvorming én voor de uitvoering van het statistisch onderzoek met resampling methodes. Maar in dit artikel beperken we ons tot bovenstaande voorbeelden. Het moge duidelijk zijn dat formele kansrekening niet aan de orde komt. Het kansbegrip zit verweven in de (gesimuleerde) kansverdeling, met de variabiliteit als belangrijk kenmerk.

We zitten daarmee goed in de internationale trend van het statistiek onderwijs.

Dat kan blijken uit het GAISE report: *Pre-college statistics education should emphasize the ways probability is used in statistical thinking; an intuitive grasp of probability will suffice at these levels.* Onder het artikel staan nog meer referenties.

**Over VUStat**

Veel van de oude mogelijkheden van de windows versie van VUStat zijn behouden.

De apps versie van VUStat werkt op alle browsers en heeft vergelijkbare mogelijkheden voor wat betreft data-analyse (voor Beschrijvende Statistiek): kansberekeningen met allerlei typen verdelingen kunnen grafisch worden aangepakt; tekstbestanden kunnen worden ingelezen, en in allerlei vormen getoond, data bestanden bewaard en aangepast, uitsplitsingen naar één van de variabelen gemaakt, kruistabellen gemaakt en geanalyseerd, en nog veel meer.

Daarnaast zijn er nu apps met mogelijkheden om te simuleren.

Resampling en vooral de bootstrap is te gebruiken als een toegankelijke weg naar schatten en toetsen. In de app Data analyse is het op verschillende plekken mogelijk om direct data door te sturen naar een van de resampling apps, waar dan verdere analyse van de steekproef kan plaatsvinden.

Daarnaast is er nog een aantal apps die de breedheid van het onderwerp statistiek laten zien. Zo kun je de wetten van de grote en de kleine aantallen onderzoeken met pakkende voorbeelden. Dat is ook het geval met het verschijnsel van de regressie naar het midden.

Referenties.

Docenten ingang lespakket: http://www.toevalindegreep/doc

Internationale trend: https://askgoodquestions.blog/

 https://teaching.statistics-is-awesome.org/

 https://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12\_Full.pdf

 https://www.statisticsteacher.org/

 https://artofstat.com/

Van Brederode & Meeter:

https://www.scienceguide.nl/2020/05/hoe-statistiek-het-schoolexamen-verdacht-maakte/