Foto: Andrey Popov

Consistente schattingen voor categoriale data gebaseerd op een mix van administratieve databronnen en enquêtes

Officiële statistieken moeten aan allerlei eisen voldoen en van zo hoog mogelijke kwaliteit zijn. Zo moeten meetfouten zo veel mogelijk worden opgespoord en verbeterd. Ook moeten de statistieken consistent zijn. In Boeschoten (2019) wordt een methode geïntroduceerd die de kwaliteitseisen simultaan aanpakt voor categoriale data. Boeschoten introduceerde in haar proefschrift een methode die de drie hierboven besproken problemen simultaan aanpakt specifiek voor categoriale data. Ze doet dit door verschillende bestaande methoden op een nieuwe manier met elkaar te combineren: enerzijds de *latenteklasseanalyse* en anderzijds *multipele imputatie*.

Laura Boeschoten

Bij het produceren van officiële statistieken maakt het Centraal Bureau voor de Statistiek (CBS) zoveel mogelijk gebruik van bestaande administratieve bronnen. Soms omvatten die de gehele onderzoekspopulatie. Zo bevat de Basisregistratie Personen (BRP) alle inwoners van Nederland. Soms slechts een deel van de populatie. Zo beheeft de Dienst Uitvoering Onderwijs (DUO) geen gegevens over diploma's behaald in het buitenland of aan particuliere instellingen, en is het ook voor personen die langer geleden een opleiding hebben gevolgd minder compleet. Als er interesse is in een statistiek over een onderwerp dat helemaal niet in administratieve bronnen wordt gemeten, dan wordt de informatie verkregen door uit de populatie een steekproef te nemen en een enquête af te nemen. Over het algemeen bevatten al deze datasets unieke *identifiers* waardoor deze bronnen op persoonsniveau aan elkaar kunnen worden gekoppeld.

Meetfouten

Data verkregen door middel van enquêtes kunnen allerlei soorten fouten kunnen bevatten. Van een fout spreken we wanneer de echte score van een eenheid (meestal een persoon of een bedrijf) op een bepaalde variabele niet overeenkomt met de score die is ingevuld in de enquête. In mijn proefschrift bespreek ik alleen meetfouten in het algemeen en ga ik niet verder in op het onderscheid tussen de verschillende soorten fouten zoals gedefinieerd door Groves et al. (2009). Minder bekend is dat administratieve data evengoed meetfouten kunnen bevatten. Bijvoorbeeld doordat iemand fouten maakt tijdens het invullen van het register, wanneer iemand vergeet de informatie in het register te updaten of wanneer iemand fouten maakt bij het ophalen van de informatie uit het register.

Ongeacht met wat voor type fouten we te maken hebben en waar in de data ze precies zitten, kunnen fouten altijd problematisch zijn en zorgen voor vertekening bij het uitvoeren van statistische analyses wanneer er niets aan wordt gedaan. Soms zijn fouten willekeurig. Wanneer personen bijvoorbeeld hun gewicht in moeten vullen in een enquête en dat niet precies weten, zal de ene persoon een hoger gewicht invullen en de andere persoon een lager gewicht, en het verschil tussen het *echte* gewicht en het *ingevulde* gewicht is de ene keer groter en de andere keer lager. Al deze verschillen zorgen bij elkaar dan voor al voor extra variantie (oftewel spreiding) en wanneer we gewicht in relatie tot een andere variabele onderzoeken, zoals bijvoorbeeld geslacht, dan zal de geschatte relatie als gevolg van deze extra spreiding door meetfout vooral minder sterk zijn. Wanneer fouten structureel dezelfde kant op wijzen, bijvoorbeeld wanneer verschillende personen altijd een lager gewicht invullen dan ze daadwerkelijk hebben, dan kunnen de effecten in sommige gevallen zelfs sterker worden. Behalve deze problematische *vertekening*, zorgen meetfouten bij het CBS nog voor twee andere problemen. Het eerste probleem heeft te maken met *onmogelijke combinaties van scores* op verschillende variabelen, en is eerder besproken door Sander Scholtus in zijn artikel in *STAtOR* (Scholtus, 2018). Het tweede probleem gaat over dat de verschillende statistieken die het CBS produceert (in de praktijk gaat dit vaak over grote hoeveelheden kruistabellen) allemaal *consistent* moeten zijn. Dit wil zeggen dat wanneer een kruistabel tussen de variabelen 'opleidingsniveau × geslacht × regio' wordt geproduceerd, en daarnaast ook een kruistabel 'opleidingsniveau × geslacht × burgerlijke staat', dat, bijvoorbeeld, het totaal aantal hoogopgeleide mannen in beide kruistabellen exact gelijk aan elkaar moet zijn. Dit probleem is eerder en uitgebreider besproken in *STAtOR* door Jacco Daalmans (Daalmans, 2019).

In mijn proefschrift introduceer ik een methode die de drie hierboven besproken problemen simultaan aanpakt specifiek voor categoriale data. Ik doe dit door verschillende bestaande methoden op een nieuwe manier met elkaar te combineren: enerzijds de *latenteklasseanalyse* en anderzijds *multipele imputatie*.

Latenteklassemodellen

Ten eerste maak ik gebruik van zogenaamde latenteklassemodellen. Latenteklassemodellen worden normaal gebruikt om (categoriale) variabelen te meten die je niet goed direct kunt meten (McCutcheon, 1987). Je kunt hier bijvoorbeeld denken aan persoonlijkheidstypen of type eetstoornissen. In dat laatste voorbeeld wordt er niet direct aan personen gevraagd 'welk type eetstoornis heb je?'. In plaats daarvan wordt er gevraagd naar een aantal zaken die indicatoren kunnen zijn voor bepaalde typen eetstoornissen, zoals overgewicht, overgeven of overmatig sporten. Een bepaalde combinatie van antwoorden op deze vragen kan een indicatie zijn voor een bepaald type eetstoornis. Bij het schatten van het model worden de relaties tussen de indicatoren en de latente variabelen bepaald.

Nadat zo'n model is geschat, geeft het ons ten eerste informatie over wat de kans is dat iemand een bepaalde score heeft op een indicator-variabele gegeven de latente klasse (bijvoorbeeld de kans dat iemand overgeeft gegeven dat de eetstoornis anorexia is), de *conditionele kansen*. Ten tweede geeft het ons informatie over wat de kans is dat iemand tot een bepaalde latente klasse behoort gegeven de scores op de indicatoren (bijvoorbeeld de kans dat iemand anorexia heeft gegeven dat hij geen overgewicht

heeft, overgeeft maar niet sport), de *posterior kansen*.

In mijn onderzoek, gebruik ik de latenteklassemodellen op een andere manier. Ik gebruik namelijk variabelen die hetzelfde meten, maar afkomstig zijn van de verschillende bronnen die we op eenheidsniveau aan elkaar hebben kunnen koppelen (Biemer, 2011). De variabelen kunnen dus zowel afkomstig zijn van administratieve bronnen als van enquêtes en kunnen zowel de gehele populatie als een deel van de populatie bevatten. Het zou bijvoorbeeld kunnen dat we drie variabelen hebben die *burgerlijke staat* meten, waarvan er twee afkomstig zijn uit administratieve bronnen en een afkomstig is uit een enquête. Deze variabelen worden gebruikt als indicatoren van een latente variabele die de *echte* burgerlijke staat meet. Om deze methode te kunnen toepassen moet er wel aan een aantal voorwaarden worden voldaan. In deze specifieke situatie is vooral een belangrijke voorwaarde dat de indicatoren onafhankelijk van elkaar gemeten moeten zijn. Wanneer een administratieve bron ook *burgerlijke staat* meet, maar deze informatie heeft verkregen uit het BRP, dan zijn deze metingen niet onafhankelijk en dus ook niet bruikbaar voor deze methode.

In het latenteklassemodel kunnen ook andere variabelen (covariaten) worden meegenomen. Dit is belangrijk, want wanneer het CBS bijvoorbeeld de kruistabel '*burgerlijke staat* × *geslacht*' moet produceren, dan moet de variabele *geslacht* worden meegenomen in het latenteklassemodel dat de *echte* scores van *burgerlijke staat* meet, omdat anders de uitkomsten van deze kruistabel vertekening kunnen bevatten. Daarnaast kunnen we in het latenteklassemodel ook restricties specificeren. Zo kunnen we ervoor zorgen dat bepaalde combinaties van scores die in de praktijk onmogelijk zijn, ook niet voor kunnen komen nadat we een specifieke score voor *burgerlijke staat* hebben toegewezen na de toepassing van het latenteklassemodel. Bijvoorbeeld, de combinatie '*burgerlijke staat* = gehuwd' en '*leeftijd*=jonger dan 5 jaar' is bij wet verboden en daarom niet mogelijk (als gevolg van meetfouten is het in de praktijk natuurlijk wel mogelijk dat deze combinatie van scores wordt geobserveerd in de data). Het zou dan ook onwenselijk zijn dat deze combinatie na de toepassing van het latenteklassemodel wel voor zou komen. Door middel van restricties op het model kunnen we daar voor zorgen.

Multipele imputatie

Een manier om onderling consistente tabellen te krijgen is door de kansen uit het latenteklassemodel te gebruiken om per eenheid een score toe te wijzen, oftewel een imputatie van de latente variabele *echte burgerlijke staat* te genereren. Daarnaast willen we de geproduceerde statistieken van variantieschattingen voorzien. De statistieken worden geproduceerd door gebruik te maken van de geïmputeerde data, maar het is wel wenselijk dat de variantieschattingen de onzekerheid bevat die wordt veroorzaakt door de ontbrekende en conflicterende waarden in de oorspronkelijke databronnen. Om dit mee te kunnen nemen maken we gebruik van multipele imputatie (Rubin, 1987), een methode die doorgaans wordt gebruikt voor ontbrekende waarden in variabelen.

Toepassingen en uitbreidingen

In mijn proefschrift wordt de hierboven beschreven methode geïntroduceerd en op verschillende manieren uitgebreid. Zo wordt het aantal ernstige verkeersgewonden per voertuigtype geschat op basis van data afkomstig van politie en ziekenhuizen en wordt het aantal werkenden op meerdere tijdspunten geschat op basis van een gecombineerde dataset uit Italië. Daarnaast wordt onderzocht of de methode gebruikt kan worden om consistente volkstellingstabellen te genereren en wordt de methode uitgebreid zodat covariaten ook op een later tijdstip aan het model toegevoegd kunnen worden.

Literatuur

Boeschoten, L. (2019). *Consistent estimates for categorical data based on a mix of administrative data sources and surveys*. Doctoral thesis. Tilburg University, Tilburg. Retrieved from: https://pure.uvt.nl/ws/portalfiles/portal/31415117/Boeschoten_Consistent_25_10_2019.pdf.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey Methodology. Second Edition. Hoboken: Wiley.

McCutcheon, A. L. (1987). Latent Class Analysis. Sage, Newbury Park.

Biemer, P. P. (2011). Latent Class Analysis of Survey Error. Hoboken: Wiley.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Scholtus, S. (2018). Automatisch opsporen van fouten in data voor officiële statistiek. *STAtOR*, 19(3), 14–18

Daalmans, J. (2019). Automatisch corrigeren van inconsistenties in de officiële statistiek. *STAtOR*, 20(3), 9–12

Laura Boeschoten werkt als postdoctoraal onderzoeker aan de Universiteit Utrecht. In oktober 2019 is zij gepromoveerd aan de Universiteit van Tilburg op het proefschrift dat in dit artikel wordt beschreven.
E-mail: l.boeschoten@uu.nl

International Prize in Statistics

De wetenschap kent veel prijzen voor excellent werk. De bekendste en meest prestigieuze zijn ongetwijfeld de Nobelprijzen. Wiskundigen kennen dan verder de Fields Medal en de Abel Prize en de informatici hebben de Turing Award. In Nederland hebben we als een van de belangrijkste wetenschappelijke prijzen de Spinozapremie, enkele jaren geleden is die aan Aad van der Vaart toegekend. Op de website van de Koninklijke Nederlandse Akademie van Wetenschappen is verder een hele lijst van prijzen te vinden, zoals bijvoorbeeld de Heineken Prijzen en de Lorentzmedaille.

Ook onze eigen VVSOR kent prijzen toe, de Jan Hemelrijk en Willem van Zwet Awards en als belangrijkste de vijfjaarlijkse Van Dantzig Prijs die dit jaar weer aan de beurt is toegekend te worden.

De internationale statistische gemeenschap stelde enkele jaren geleden vast dat ook op hun vakgebied behoefte was aan een prijs voor zeer excellente wetenschappers. Er kwam een samenwerkingsverband tot stand tussen het International Statistical Institute, de American Statistical Association, het Institute of Mathematical Statistics, de International Biometric Society en de Royal Statistical Society. Samen hebben zij een stichting opgericht die tweejaarlijks de International Prize in Statistics toekent. De uitreiking vindt plaats tijdens het tweejaarlijks World Statistics Congress van het ISI. Aan de prijs is een geldbedrag van US$ 75.000 verbonden.

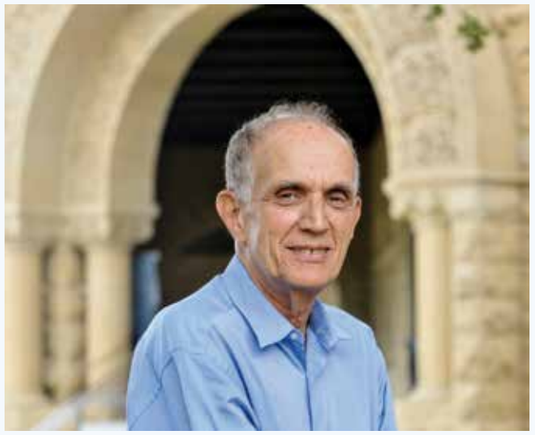
Op de website van de organisatie staat de volgende informatie over deze prijs:

- The prize will be awarded for a single work or body of work, rather than for more diffuse reasons such as "lifetime achievement." Not only should powerful and original ideas be recognized by the prize, but also contributions that lead to breakthroughs in other disciplines or works with important practical effects on the world.
- Generally, the prize will be awarded to individuals, but in some cases, groups of individuals working on similar ideas — or even teams of individuals or organizations — could be recognized.

Na de eerste toekenningen in 2017 aan Sir David R. Cox in 2019 aan Bradley Efron is het nu de beurt aan de prijs voor 2021, die tijdens het WSC-2021 in Den Haag zal worden uitgereikt. Het nominatieproces is onlangs gestart. Iedere statisticus kan een voordracht indienen, alle informatie hierover is te vinden op de genoemde website.

Voordrachten voor de International Prize in Statistics 2021 kunnen tot 15 augustus 2020 via de website worden aangemeld. Zie voor meer informatie https://statprize.org/

Sir David R. Cox


Bradley Efron