Casper Albers
Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column Casper grijpt een kans

Column van het jaar

December is een speciale maand. Het is de maand van Sinterklaas, Kerst en de Oud van Oud & Nieuw. Het is ook de tijd van het jaar voor eindejaarslijstjes. Casper rekent u voor wat er mis kan gaan in dit soort lijstjes.

Wanneer stemmen?

Eindejaarslijstjes als 'film van het jaar', 'ondernemer van het jaar' en 'speelgoed van het jaar' zijn een jaarlijks terugkerend fenomeen. Het is een beetje raar dat dat soort lijstjes al verschijnen voordat het jaar voorbij is. Bij een voetbalwedstrijd wijs je ook niet in de tachtigste minuut al de winnaar aan.

Soms lijkt dat niet zo'n probleem te zijn. Zo las ik eind september een nieuwsbericht waarin stond dat Luis Fonsi's 'Desposito' officieel verkozen was tot zomerhit van het jaar. (Welke instantie de bevoegdheid heeft zulke 'officiële' besluiten te maken, weet ik niet, maar dat is geen discussie voor een wiskundetijdschrift.) Het klinkt acceptabel om te bepalen wat de zomerhit was zodra het herfst is. Het wordt al iets discutabeler als je weet dat dat nummer al op 13 januari (in de winter dus) werd uitgebracht en in mei (in de lente dus) al op de nummer 1-positie kwam. Een ander voorbeeld waar het wel kan: op 23 oktober werd Lieke Martens door de FIFA tot beste voetbalster van de wereld gekozen. Aangezien de belangrijkste voetbalwedstrijden toen al achter de rug waren, was dat een veilige timing.

Vaak is het echter wel een probleem. Zo werd op 28 september Ingrid de Graaf van Aegon tot 'topvrouw van het jaar' gekozen [1]. Met nog 26% van het jaar te gaan een gewaagde timing, al zal het volgens de vakjury vast niet waarschijnlijk zijn geweest dat De Graaf in de laatste maanden iets doet waardoor ze opeens geen goed bestuurder meer is.

Anders is het bij verkiezingen waarbij je pas aan het eind van het jaar echt weet wat de kandidaten zijn. Zo organiseert de Van Dale jaarlijks de 'Woord van het Jaar'-verkiezing [2]. Hier bepaalt de bezoeker van de website van de Van Dale welk woord deze prestigieuze trofee op de schoorsteenmantel mag zetten. Klein probleem: de verkiezingsperiode begon begin november. Om in aanmerking te komen, moest een woord dus ergens tussen januari en oktober ontstaan zijn. Als we er, bij gebrek aan meer kennis, van uitgaan dat nieuwe woorden op willekeurige momenten in het jaar hun intrede doen, betekent dat dus dat de kans dat het beste woord van het jaar pas in november of december ontstaat ongeveer $\frac{2}{12} = \frac{1}{6}$ is. Gemiddeld eens in de zes jaar is het 'Woord van het Jaar' dus eigenlijk op z'n best het 'Een-na-beste Woord van het Jaar'.

Omdat de Van Dale niet alleen de winnaar maar de gehele top 3 bekendmaakt, wordt de kans nog groter dat daar een woord uit november/december in mist. Die kans is $1-(1-\frac{1}{6})^3 = \frac{91}{216} \approx 42\%$. Zou de Van Dale de top 4 bekendmaken, dan is de kans groter dat dit verkeerd is dan dat dit juist is.

Meerdere verkiezingen

Schrale troost is dat ook het Genootschap Onze Taal (GOT) jaarlijks bekendmaakt wat het Woord van het Jaar is. Hoewel ook deze prijs voor het einde van het jaar wordt bepaald, is het interessant om te zien dat bij beide instanties vaak een ander woord de trofee in de wacht sleept. Vorig jaar won bij Van Dale het woord *treitervlogger* met 35% van de stemmen [3]. Bij het GOT won het woord *brexit* en kreeg treitervlogger slechts 8% van de stemmen [4]. Dit verschil kan door twee oorzaken komen: de populatie van potentiële Van Dale-stemmers wijkt af van de populatie van potentiële GOT-stemmers, of er was toeval. Omdat dit een statistiekcolumn is, gaan we dit toetsen en we beperken ons voor het gemak tot het woord treitervlogger. Formeel is de toetssituatie als volgt:

Laat π_1 en π_2 de proporties treitervlogger-liefhebbers zijn onder de populaties van potentiële stemmers bij Van Dale (1) en het Genootschap Onze Taal (2), en laat $p_1 = 0{,}35$ en $p_2 = 0{,}08$ de geobserveerde proporties zijn. De nulhypothese $\mathrm{H}_0\colon \pi_1 = \pi_2$ versus

tweezijdig alternatief kan getoetst worden met een Z-toets met

$$Z = \frac{p_1 - p_2}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

waarbij $p^* = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$ een gewogen gemiddelde van beide proporties is. Als $|Z| > 1{,}96$, dan verwerpen we H_0 (indien we op een niveau van 5% toetsen). Een interessante complicatie is dat we hier niet weten hoe groot n_1 en n_2 zijn: het persbericht van Van Dale meldt alleen dat $n_1 > 100000$ en het bericht van het GOT zegt niets over het aantal stemmen. We weten alleen dat het percentage p_1 afgerond 8% is. Dit kan voor $n_2 \in \{12, 13, 24, 25, 26, 36, \ldots\}$. Voor zowel n_1 als n_2 zijn wel ondergrenzen te vinden (waarbij het wat onrealistisch is om aan te nemen dat $n_2 = 12$) en hoe groter n_1 en n_2, hoe groter Z. De p-waarde bij $n_1 = 100000$ en $n_2 = 12$ is gelijk aan $p = 0{,}049895$, nét kleiner dan de 5%-grens. Bij een $n_2 > 100$ geldt al dat $p < 10^{-7}$, dus we hebben hier wel een significant verschil te pakken: Van Dale-stemmers stemmen echt anders dan GOT-stemmers.

Hoe de stemmen te tellen

Maar zo'n verkiezing van X van het jaar is zo makkelijk nog niet. Hoe bepaal je de winnaar? Zowel Van Dale als GOT vragen stemmers om hun nummer 1-keuze, en kijken welk woord het vaakst gestemd wordt: dat woord wint. (En dan maar hopen dat november en december saaie maanden zijn qua nieuw vocabulaire.) Het kan ook anders: je kan bijvoorbeeld iedereen vragen om een top 3 in te sturen en dan turven wat het meest genoemd wordt. In Tabel 1 staat een virtueel voorbeeld: Alice, Bob en Carol hebben een rangschikking gemaakt van vier woorden (de top 3 van Van Dale en de winnaar bij GOT). Zou je alleen de nummers 1 tellen, dan wint Brexit van Trumpisme, maar zou je top 3's tellen, dan zijn de verhoudingen omgedraaid: Trumpisme staat in elke top 3, Brexit maar in eentje. De gemiddelde ranking van Trumpisme, $2\frac{1}{3}$, is ook beter dan die van Brexit, 3.

Het maakt dus nogal uit of je alleen de kopposities meet of ook de nummers $2, 3, \ldots, I$ voor een zekere I. Een omslachtige manier om dit op te lossen is om meerdere stemrondes te houden: als je N kandidaten hebt voor de hoofdprijs, kan je $N-1$ stemrondes houden. Bij elke ronde laat je dan de kandidaat met de minste steun wegvallen. Dit is vooral nuttig wanneer kandidaten in elkaars vaarwater zitten: als bijvoorbeeld bij presidentsverkiezingen een linkse kandidaat wegvalt, zal de steun voor een andere linkse kandidaat doorgaans toenemen. Een andere manier om via meerdere stemrondes de winnaar te bepalen is via een knock-outsysteem waarbij telkens verkiezingen zijn tussen twee kandidaten en de winnaar doorgaat naar de volgende ronde. Ook dan heb je aan $N-1$ stemrondes voldoende.

Woord	Alice	Bob	Carol
Treitervlogger	1	3	3
Pokémonterreur	2	1	4
Trumpisme	3	2	2
Brexit	4	4	1

Tabel 1 Rankings gegeven aan vier woorden uit 2016.

Condorcet-paradox

Bij stemprocedures gebaseerd op rangschikkingen kan een interessante paradox optreden, vernoemd naar de markies van Condorcet die deze paradox al in de achttiende eeuw omschreef. Uit Tabel 1 is te af te lezen dat een (tweederde) meerderheid van de stemmers Treitervlogger een beter woord vindt dan Pokémonterreur. Tevens vindt een (tweederde) meerderheid dat Pokémonterreur een beter woord is dan Trumpisme. Intuïtief zou je verwachten dat er dus een meerderheid is die Treitervlogger verkiest boven Trumpisme. Dit is echter niet het geval, er zit een soort lus in de voorkeuren. Welk van deze drie woorden ook tot winnaar wordt uitgeroepen: er is een ander woord dat meer steun van de stemmers had!

Zo'n lus kan niet altijd optreden. Als 100% van de stemmers A prefereert boven B en B boven C, dan is het onmogelijk dat toch een meerderheid C boven A zet. Een voorwaarde voor het hebben van zo'n lus is het volgende: laat p_1 de proportie stemmers zijn die A prefereert boven B, p_2 de proportie met B boven C en p_3 de proportie met C boven A. Door de labels A, B en C te permuteren, is het altijd mogelijk om $p_1 \geq \frac{1}{2}$ en $p_2 \geq \frac{1}{2}$ te veronderstellen. Er kan afgeleid worden [5] dat $p_3 \leq 2 - p_1 - p_2$. De paradox treedt op als $p_2 > \frac{1}{2}$ (en de '$\geq$'-relaties voor p_1 en p_2 strikte '$>$'-relaties zijn). Dit impliceert de voorwaarde $p_1 + p_2 < \frac{3}{2}$.

Conclusie

Alles samenvattend: zo makkelijk is het nog niet om een jaarverkiezing te houden. Hoe meer je er over puzzelt, hoe meer puzzels boven water komen. Een eenvoudige basisregel is echter wel om een verkiezing pas te houden nádat het jaar afgelopen is. Daar zit geen hogere wiskunde achter en daar is al winst te halen.

Referenties

1 https://www.nrc.nl/nieuws/2017/09/28/ingrid-de-graaf-is-topvrouw-van-het-jaar-a1575361

2 http://woordvanhetjaar.vandale.nl

3 http://www.vandale.nl/woord-van-het-jaar-2016-treitervlogger

4 https://onzetaal.nl/nieuws-en-dossiers/weblog/brexit-woord-van-het-jaar-bij-onze-taal/

5 C.L. Silver, The voting paradox, *The Mathematical Gazette* 76(477) (1992), 387–388. http://www.jstor.org/stable/pdf/3618386.pdf