Peilingen: de logica van de onzekerheid

Simon van der Salm

Presidentsverkiezingen in de USA

Vóór de Amerikaanse presidentsverkiezingen in november 2016 stond *Hillary Clinton* steeds ruim voor in de opiniepeilingen. In oktober 2016, een maand voor de Amerikaanse verkiezingen, gaven de media Clinton een zeer grote kans (er werden getallen genoemd van meer dan 90%, maar het is onduidelijk wat ze daarmee bedoelen) op het verslaan van Trump, maar een paar weken later was er massaal ongeloof toen bekend werd dat Clinton de verkiezing tot president van de USA bleek te hebben verloren.

Velen vroegen zich af, hoe het mogelijk was dat de peilingen er zo ver naast zaten. En die vraag wordt ook nu weer gesteld. De uitslag van de presidentsverkiezingen in november 2020 bleek flink af te wijken van de waarden in de peilingen die eerder (vanaf het voorjaar tot zelfs tot vlak voor de verkiezingen) werden waargenomen. Nu is Biden de winnaar. Zie figuur 1. Duiden die afwijkingen op wetenschappelijk onvermogen van de statistici die de peilingen uitvoeren, of is er iets anders, en vooral iets meer, aan de hand? Gaan de peilers niet uit van veel te eenvoudige, en dus ineffectieve modellen?

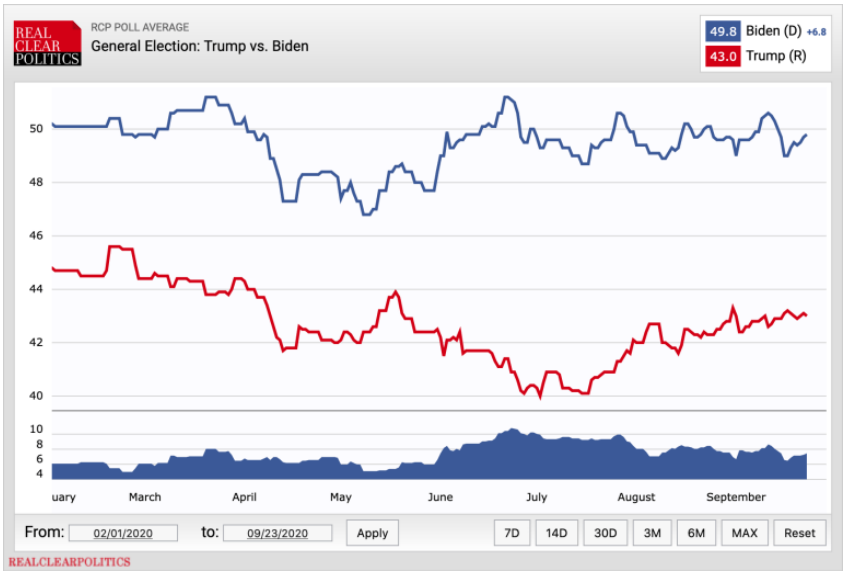

Fig. 1. Voorsprong van Biden op Trump. Op 22 september had Biden een voorsprong van 6,6%-punten op Trump. Bron: RealClearPolitics/. Op de verkiezingsdag was daar nog (maar) 3,2%-punten van over.

Onzekerheidsinterval volgens Wald

Om een eerste idee over een mogelijk antwoord te krijgen, kijken we naar het eenvoudigste wiskundige model dat de basis vormt van opiniepeilingen.

Voordat verkiezingen worden gehouden willen politieke partijen, nieuwsmedia en consumenten graag beschikken over een schatting van de populatieproportie π van kiezers op de politicus X. Uit de populatie potentiële kiezers neemt een deskundige peiler een steekproef van grootte n. Gebruikelijk is een waarde rond de 1.000. Vervolgens bepaalt hij de steekproefproportie p van personen die voor kandidaat X zeggen te zullen gaan stemmen.

De steekproefproportie p is een *schatting* van de 'werkelijke' populatieproportie π. Die schatting is vanzelfsprekend niet oneindig nauwkeurig: door allerlei toevalsfactoren zal de steekproefproportie p meer of minder van de populatieproportie π afwijken. Er is sprake van een onvermijdelijke *foutmarge*. De foutmarge is de halve breedte van een onzekerheidsinterval en gemakkelijk te berekenen. Daarvoor bestaan zelfs rekenlinialen. Zie figuur 2.

Om te begrijpen wat een onzekerheidsinterval (betrouwbaarheidsinterval) is, bestuderen we een eenvoudig voorbeeld. Een vaas bevat 10.000 blauwe en 10.000 rode knikkers, goed door elkaar gehusseld.

Uit de vaas nemen we een *aselecte steekproef* (= een steekproef, waarin iedere knikker met dezelfde kans terecht kan komen) van grootte n = 1.000 en bepalen het percentage p van blauwe knikkers daarin.

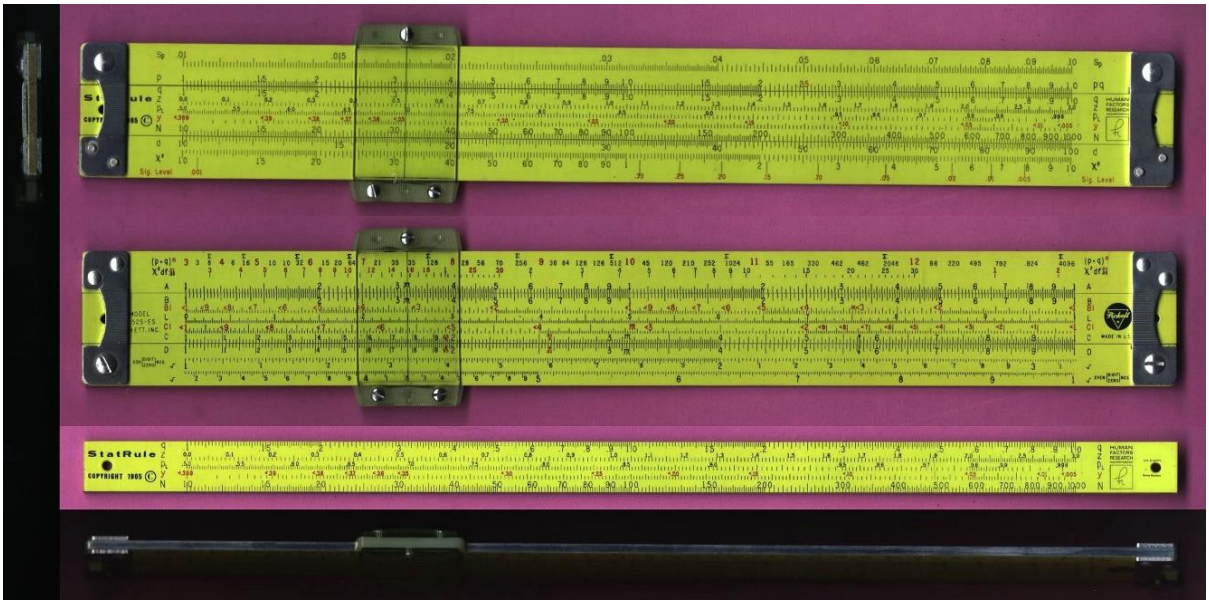

Fig. 2. De Pickett N 525-ES StatRule. Bron: https://www.sliderulemuseum.com/Pickett.htm. Met deze liniaal kunnen Wald-onzekerheidsintervallen worden berekend. Zie voor beschrijving en uitleg [2].

Wald's equation for the normal approximation of the binomial uncertainty interval is:

$$\pi = p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

π: population proportion

n: sample size

p: sample proportion

Restriction: $np > 5$ and $n(1-p) > 5$

$1-\alpha$: the confidence level

$z_{1-\frac{\alpha}{2}}$: the z-value of the confidence level

Formule-box met de formule van Wald (1943). Opmerkelijk is dat Pierre-Simon Laplace deze formule al vermeldt in zijn Théorie analytique des probabilités (1812).

Bijvoorbeeld p = 48%. Bij waarden van p, dicht bij 50%, is de 95%-meetonzekerheid (de foutmarge bij de veel gehanteerde betrouwbaarheid van $\gamma = 1 - \alpha$ = 95%), volgens de formule van Wald in de formule-box, ongeveer 3,1%, namelijk $u \approx \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1000}}$. Het 95%-onzekerheidsinterval loopt daarom van 44,9% tot 51,1%.

Dat interval bij p = 48% *bedekt* inderdaad de populatieproportie π = 50%. We doen vervolgens de eerder getrokken knikkers terug in de vaas. In een volgende steekproef van 1.000 knikkers zouden we bijvoorbeeld p = 55% kunnen vinden, ondanks het feit dat de populatie gelijk is gebleven. Het 95%-onzekerheidsinterval loopt dan van 51,9% tot 58,1%. In dat geval bedekt het interval *niet* de populatieproportie π = 50%. Dat is geen ongewoon verschijnsel. Bij gemiddeld 1 op de 20 steekproeven (= 5% onbetrouwbaarheid) kunnen we een onzekerheidsinterval rond de steekproefproportie p verwachten dat de populatieproportie π mist. That's all-in the game en onderdeel van de logica van statistische onzekerheid! Zie referentie [4].

Op internet zijn fraaie software calculators te vinden voor het berekenen van onzekerheidsintervallen, bijvoorbeeld https://istats.shinyapps.io/ExploreCoverage/. Daarmee kunnen we met gemak bijvoorbeeld

het trekken van 500 onafhankelijke steekproeven met n = 1000 simuleren. Zie figuur 3. In deze simulatie bedekt 'toevallig' 93,4% de populatieproportie en niet de nominale 95%. Daarom mist 6,6% van de intervallen de werkelijke proportie π = 50%.

Dit laat zien dat, naast het feit dat we met een onzekerheidsinterval te maken hebben, ook de feitelijke bedekkingskans aanzienlijk kan verschillen van de nominale bedekkingskans van 95%. Zie referenties [1] en [5].

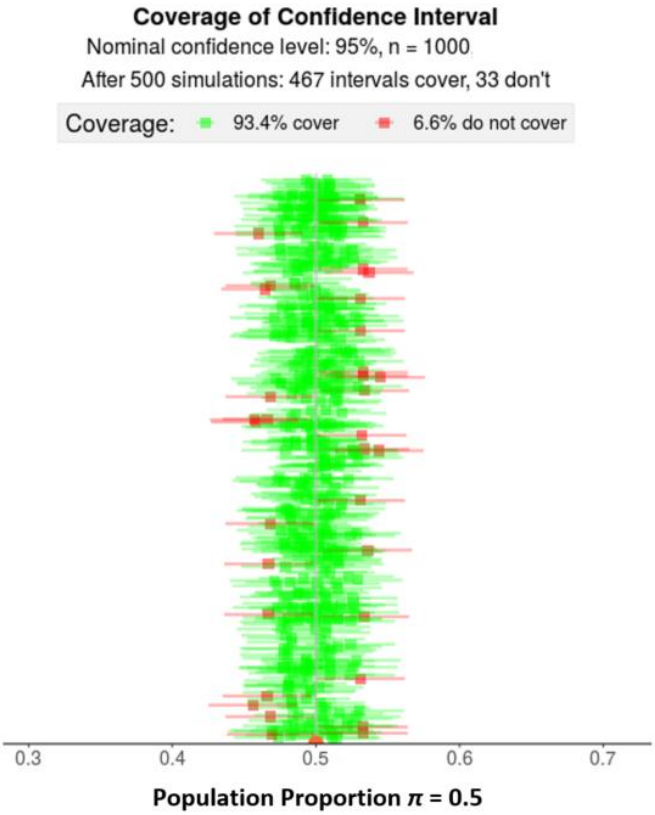

Fig. 3. 500 Wald-intervallen van 500 gesimuleerde steekproeven. De rode intervallen bedekken niet de populatieproportie π = 0,5.

Representativiteit

Als we kijken naar de formule van Wald, dan valt direct op dat we de meetonzekerheid aanzienlijk kunnen verkleinen door n flink groot te maken. Helaas neemt de foutmarge maar met $\sqrt{n}$ af, dus je moet n aanzienlijk vergroten voor een significante verlaging van de meetonzekerheid. Het verhogen van n is wellicht handig voor een eenvoudige, kunstmatige populatie, zoals in bovenstaand voorbeeld met de knikkers, maar bij de echte populaties, waar opiniepeilers mee te maken hebben, van niet altijd even consistente, menselijk kiezers, zijn er meerdere factoren die roet in het eten gooien.

Het nemen van grote steekproeven uit een echte populatie van kiezers is moeilijk en tijdrovend, en daardoor erg duur. Met voldoende (financiële) middelen is aan dat bezwaar natuurlijk betrekkelijk gemakkelijk tegemoet te komen. De moeilijkheden van het nemen van een goede steekproef uit een werkelijke populatie kiezers zitten echter primair, niet in de steekproefgrootte n, maar in de *representativiteit*. Een aantal problematische aspecten zijn:

1. Het aantal respondenten in een door de peiler uitgezochte steekproef moet voldoende groot zijn. Als het gaat om een proportie π in de Nederlandse bevolking, dan is een steekproef van n = 100 aan de kleine kant, maar een steekproef van n = 700 of meer, kan als voldoende groot beschouwd worden. Dat aantal wordt bepaald door de maximale foutmarge die de opiniepeiler wenst te accepteren. 'Grote' steekproeven hebben in de praktijk een n van 1.000 tot 1.500; het is, om allerlei redenen, niet erg zinvol om n nog groter te kiezen.

   Een vervelend (en ook toenemend) probleem bij echte steekproeven is namelijk het aantal niet-deelnemers, de *non-respons*. De verhouding deelnemers/niet-deelnemers, het *deelnemingspercentage*, moet voldoende hoog zijn. Het aantal niet-deelnemers aan een steekproef in een onderzoek kan in extreme gevallen wel 80% zijn, dus als je in zo'n geval uiteindelijk 700 deelnemers in je steekproef wilt overhouden, moet je 3.500 personen uitkiezen. En je hebt geen goed idee

waarom de 2.800 niet-deelnemers niet mee willen doen. Die kunnen om toevallige redenen niet mee doen (uitnodiging niet of te laat ontvangen, geen tijd wegens familieomstandigheden, enzovoorts). Dat is uiteindelijk niet zo erg, maar ze kunnen ook bewust niet meedoen, omdat ze bijvoorbeeld een aversie tegen de politiek hebben. Of van mening zijn dat de opiniepeiler niet onafhankelijk is. Misschien zijn de niet-deelnemers wel een bijzondere (en dus relevante) deelpopulatie van het grotere geheel, en dat zou je als peiler graag willen weten. Een hoge respons is weliswaar een noodzakelijke indicatie voor representativiteit, maar niet een voldoende. We komen daarmee op het volgende punt.

2. De compositie van de steekproef (geslacht, leeftijd, inkomensklasse, opleidingsniveau, enzovoorts) moet overeenkomen met de partionering van de populatie. Ook als de steekproef voldoende groot is en tevens het deelnemingspercentage voldoende hoog is, dan nog kan een verkeerd gecomponeerde steekproef een aanzienlijke en ongewenste vertekening in het beeld van de werkelijkheid geven. Denk bijvoorbeeld aan het geval dat de bewoners van een specifiek, voornamelijk uit villa's bestaand, stadje oververtegenwoordigd zijn in een steekproef naar politieke voorkeur. Een ander illustratief voorbeeld is het bellen van de slinkende groep mensen die in het telefoonboek staan (dat is zo'n lekker gemakkelijk te bereiken groep) en die dus een vaste telefoonlijn gebruiken. De peiler moet dan niet verbaasd zijn dat hij voornamelijk oudere personen in zijn steekproef vindt.

3. Een heel moeilijke eis is dat er geen systematische verschillen mogen bestaan tussen deelnemers en niet-deelnemers in de uitgezochte steekproef. Die eis is zo problematisch omdat over de niet-deelnemers gewoonlijk veel minder bekend is, en dat ook moeilijker te achterhalen is, dan over degenen die toezeggen wel te willen deelnemen.

Kortom, het nemen van een goede representatieve steekproef is uitermate moeilijk; in veel 'echte' gevallen, zoals bij verkiezingspeilingen, lijkt dat zelfs nagenoeg onmogelijk te zijn.

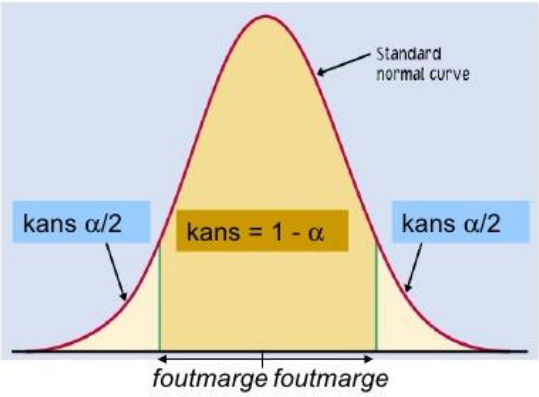

Fig. 4. De formule van Wald voor de foutmarge van een populatieproportie π is gebaseerd op een normale, dus continue benadering van de binomiale, dus discrete verdeling, waardoor er een discontinuïteitsfout optreedt. De formule in bovenstaande box is niet daarom goed bruikbaar voor kleine steekproeven en kleine of grote waarden van p. Bron: https://www.slideshare.net/fredebecker/paragraaf.

Voorspelling?

Wat velen vergeten is dat peilingen geen voorspellende waarde hebben. Men realiseert zich niet dat peilingen momentopnamen zijn, met de nadruk op het woord *moment*. En dat moment is al voorbij als de peiling eenmaal is gehouden. Daarna zijn er weer oneindig veel andere momenten, en dus vermoedelijk andere peilingsresultaten, die ieder ook weer op een bepaald moment zijn gemaakt.

Op uitzonderlijke gevallen na, is het moment dat de resultaten van een peiling bekend worden gemaakt niet hetzelfde als het afnamemoment van de peiling. In tegenstelling tot het bovenstaande, kunstmatige geval met de vaas met knikkers, kan in de tijd tussen afname en publicatie van de peiling de werkelijkheid (de voor de steekproef relevante opvattingen in de populatie) al weer veranderd zijn, bijvoorbeeld door misleidende informatie in de media, door politieke propaganda, enzovoorts. De resultaten van verkiezingspeilingen worden dus continu door toevallig en inconsistent nieuws beïnvloed en kiezers zijn

zelf ook vaak niet eenduidig. Daardoor bevatten de resultaten van alle verkiezingspeilingen een kleinere of grotere mate van inconsistentie die een voorspelling min of meer onmogelijk maken.

Gewogen groepen in de steekproef

Om een steekproef representatief te krijgen worden allerlei deelgroepen van de populatie in de steekproef *gewogen*. De steekproef wordt *gekalibreerd* op de populatie. De groep respondenten met een vaste telefoonlijn bijvoorbeeld krijgt dan een ander gewicht dan groepen in de steekproef die op een andere manier zijn benaderd, zoals via internet, of via enquêteurs die op postcode geselecteerde woningen bezoeken. Dat toekennen van allerlei gewichten lijkt indrukwekkend en objectief, maar is in werkelijkheid in meer of mindere mate subjectief.

Er kunnen ook heel andere verschijnselen een rol spelen. Aanhangers van de partij van wie de kandidaat mogelijk gaat verliezen, reageren minder op het verzoek om mee te doen aan een steekproef, terwijl aanhangers van de potentiële winnaar juist wel mee willen doen. Mensen die maatschappelijk betrokken zijn doen graag mee aan peilingen, anderen weer niet. Enzovoorts.

Kortom, peilingen zijn gebaseerd op modellen, en die modellen zijn gebouwd op (twijfelachtige) vooronderstellingen over wie in welke deelgroep in de populatie waarschijnlijk zal gaan stemmen en wie niet, en over hoeveel gewicht elk demografisch cohort moet krijgen in de steekproef van waarschijnlijke kiezers.

Fig. 5. Bron: De Volkskrant, 4 november 2020, beeld van AFP.

Inherente feilbaarheid

Ondanks dat de peilingen zijn verlies leken aan te kondigen, won Trump in 2016 van de gedoodverfde winnaar Clinton. Biden won de verkiezing in 2020, maar met een veel kleinere voorsprong dan de peilingen tot kort voor de verkiezing voorzagen. Die overwinningen benadrukken nog eens de *inherente feilbaarheid* van peilingen. De peilingen van 2016 konden bijvoorbeeld geen rekening houden met kiezers op het platteland die op het laatste moment voor Trump kozen of met Afro-Amerikanen in de sleutelstaten, wie het stemmen door de Trump-campagne onmogelijk werd gemaakt.

Peilingen hebben dus de onvermijdelijke neiging slecht of matig gekalibreerd te zijn, wat een valide resultaat in de weg staat. Bij elke peiling hoort dus het voorbehoud in gedachten gehouden te worden dat de resultaten alleen zinvol zijn als het electoraat eruit ziet zoals de peiler denkt dat het eruit zal zien op de verkiezingsdag. Peilen is dus moeilijk, want er is niet één bevolking met constante samenstelling, en met constante overtuigingen, die bestaat tot aan de verkiezingen. Er existeert dus niet zo iets als een

constante populatieproportie π. Die proportie is een min of meer heftig bewegende functie van de tijd, dus, $\pi = \pi(t)$, en het lijkt erop, onmogelijk nauwkeurig waar te nemen.

Fig. 6. De Hongaars-Joodse wiskundige en statisticus Abraham Wald (1902 – 1950). Zie [3] voor een biografie.

Consumentenwaarschuwing

Hoewel dat nog altijd niet de gewoonte is, horen ook in de publieksmedia uitslagen van opiniepeilingen gepaard te gaan met een *consumentenwaarschuwing*. Dat wil zeggen, de vermelding van de foutmarge, die de gebruiker van een peiling met de neus op het onontkoombare feit drukt dat de enquête een aanzienlijke meetonzekerheid bevat. Die foutmarge is niet een bekentenis dat de opiniepeiling verkeerd is, integendeel, de foutmarge is een schatting voor het bereik van mogelijke *toevallige* en andere *onvermijdbare* onzekerheden in de peiling, die de consument dient te kennen.

En een dergelijke waarschuwing kan voorkomen dat de consument verkeerde conclusies trekt. Als bijvoorbeeld voor een kandidaat een steekproefproportie van 51% gevonden wordt, met een foutmarge van 3%-punten, dan zal *waarschijnlijk* de 'echte' proportie π, *op dat moment*, waarschijnlijk ergens liggen in het interval van 48% tot 54%. Maar, dat is absoluut niet zeker! De betrouwbaarheid van de peiling is immers geen 100%. Als er maar twee kandidaten zijn, zoals bij de presidentsverkiezingen in de USA, dan geldt dezelfde foutmarge voor de andere kandidaat, voor wie de proportie 49% is waargenomen in de steekproef. Zijn 'werkelijke' proportie π zal dus, op dat moment, *vermoedelijk* ergens tussen 46% en 52% liggen (met een betrouwbaarheid van 95%). De intervallen overlappen elkaar. Dat betekent dat de steekproef helemaal geen significant verschil waarneemt tussen de twee kandidaten. Dus de kandidaat met 51% van de stemmen in de steekproef alvast tot winnaar uitroepen, is voorbarig en onterecht.

Recent onderzoek suggereert dat de berekende foutmarge slechts het kleinere deel van de potentiële steekproefproblemen voor zijn rekening neemt. Sommige respondenten begrijpen bijvoorbeeld niet wat de onderzoeker vraagt, andere zijn niet bereid eerlijk te zeggen welke kandidaat ze gaan steunen op de verkiezingsdag, en dergelijke. Kortom, er zijn veel andere mogelijke fouten dan de wiskundige, die de meetonzekerheid vergroten.

Fig. 7. Pierre-Simon Laplace (1749 -1827) publiceerde over mathematische fysica, hemelmechanica en veel op het gebied van de waarschijnlijkheidsrekening. Laplace kende al de formule van Wald.

Er bestaat een simpele, maar twijfelachtige en subjectieve oplossing voor het kwantificeren van zulke 'sociologische' onzekerheden: stel de totale meetonzekerheid van het steekproefresultaat gelijk aan het dubbele van de oorspronkelijk berekende, mathematische foutmarge, die dat type onzekerheden uit de aard der zaak niet mee kan nemen. Dat zou in de praktijk een foutmarge bij een grote steekproef in de orde van grootte van 5%- of 6%-punten betekenen en de vraag is, wat bij een dergelijke (weliswaar eerlijkere) onnauwkeurigheid dan nog de waarde van de peiling is.

Een technologische oplossing?

Door al die nagenoeg onoplosbare problemen met de representativiteit van steekproeven zijn peilers op zoek gegaan naar bruikbare alternatieven. Een alternatief is het geautomatiseerd doorvlooien van sociale media. Computers met kunstmatige intelligentie kunnen data-analyses uitvoeren op een schaal (steekproeven van vele tien- of honderdduizenden 'respondenten') en detaillering (bijvoorbeeld de analyse van het taalgebruik) die onmogelijk te bereiken is voor de klassieke peilingen met (maar) zo'n 1.000 deelnemers. Zulke zelflerende computerprogramma's (algoritmen) beschouwen de groep gebruikers van sociale media als een soort van *always-on* focusgroep om een meer real-time inzicht in trends te krijgen.

Conclusie

Het trekken van een volledig representatieve steekproef uit een 'echte' populatie, van bijvoorbeeld kiezers, is zeer moeilijk, zo niet onmogelijk, waardoor de meetonzekerheid van het peilingsresultaat aanzienlijk groter is dan aanvankelijk wiskundig/statistisch is berekend. Bovendien is het resultaat van de peiling een momentopname, zonder voorspellende waarde.

Referenties

[1] Zie het lemma *Coverage Probability* in Wikipedia.
[2] Salm, Simon A.M. van der, *The Pickett N-525 StatRule Statistical Scales*, Journal of the Oughtred Society, Volume 26, Number 1, Spring 2017.
[3] Biografie van Abraham Wald: https://en.wikipedia.org/wiki/Abraham_Wald.
[4] De term *logica van onzekerheid* is bedacht door de wiskundige/statisticus De Finetti. Zie de biografie https://nl.wikipedia.org/wiki/Bruno_de_Finetti.
[5] Over de meest gebruikte varianten van het binomiale betrouwbaarheidsinterval voor een proportie en de bijbehorende bedekkingskansen, is heel veel geschreven en te vinden op internet. Een goede bron is https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval.

Barbie en de Texas Instruments SR 50 — Simon van der Salm

Stadsmuseum Almelo

Niet alleen de, (meest) mannelijke, verzamelaars van rekenlinialen hebben een verslavende gewoonte, ook vrouwelijke verzamelaars kunnen soms geen maat houden. De Duitse *Bettina Dorfmann*, de grootste verzamelaarster ter wereld, heeft maar liefst 18.000 barbiepoppen…eh…verschillende. Afgelopen zomer had het Stadsmuseum in Almelo een betrekkelijk kleine Barbie-expo met *maar* 600 poppen uit haar collectie, maar toch al zo veel dat die expositie de aandacht van de journaals en andere landelijke media trok. Zie referentie [1].

Soms komen rekeninstrumenten op je weg, op plaatsen waar je die absoluut niet verwacht. Toen ik afgelopen zomer met mijn vrouw de Barbie-Expo in het Stadsmuseum van Almelo bezocht, zag ik in een vitrine van de vaste tentoonstelling over de geschiedenis van de Twentse stad, een Texas Instruments