

Rianne de Heide
 Afdeling Wiskunde
 Vrije Universiteit Amsterdam
 r.de.heide@vu.nl

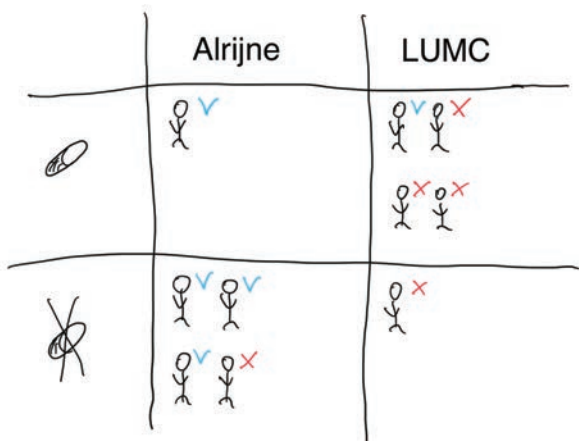


Column Rianne grijpt haar kans

Statistiek als kookboek

Rianne de Heide zal op regelmatige basis in dit blad een column schrijven over een actueel statistisch onderwerp.

Stel we willen weten of een nieuwe medicijn beter werkt dan de huidige standaard. We testen het in twee ziekenhuizen: het Alrijne in Leiderdorp en het LUMC in Leiden. Aan sommige patiënten die binnenkomen geven we het nieuwe medicijn, en aan andere het oude. In werkelijkheid gaat het eerder om tientallen of honderden, maar om het voorbeeld eenvoudig te houden, houden we het hier zonder verlies van algemeenheid klein. Stel de aantallen zijn als in Figuur 1. We zien hier dat één persoon in het Alrijne het nieuwe medicijn heeft gekregen, en vier personen in het LUMC. In het Alrijne hebben vier personen het oude medicijn gekregen, en één in het LUMC. Stel nu dat de uitkomsten zijn zoals in de figuur: de persoon in Alrijne met het nieuwe medicijn herstelt, in het LUMC herstelt er één en drie niet. Van de personen in het Alrijne met het oude medicijn herstellen er drie, de vierde niet, en ook degene in het LUMC herstelt niet.



Figuur 1

Laten we kijken naar het herstelpercentage voor de vier subgroepen:

- Alrijne + Nieuw: 100%
- Alrijne + Oud: 75%

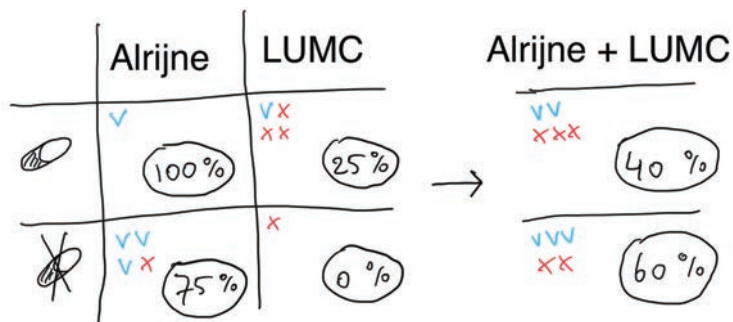
Het lijkt dat in de groep patiënten in het Alrijne, het nieuwe medicijn een verbetering brengt in het herstelpercentage.

- LUMC + Nieuw: 25%
- LUMC + Oud: 0%

Ook in de groep patiënten in het LUMC lijkt het nieuwe medicijn een goed resultaat op te leveren. Maar wat gebeurt er, als we de patiënten in beide ziekenhuizen op een hoop gooien (zie Figuur 2), en dan kijken naar het herstelpercentage?

- Nieuw: 40%
- Oud: 60%

Huh? Nu lijkt het nieuwe medicijn juist schadelijk! Wat is hier aan de hand?



Figuur 2

Simpsons paradox

Dit is een voorbeeld van Simpsons paradox. En dit is niet iets wat kan worden opgelost met wiskunde en statistiek. We moeten ons verdiepen in de achterliggende praktijk om de juiste conclusie te kunnen trekken. Het kan bijvoorbeeld zo zijn dat de ziekere patiënten in het LUMC terecht komen, en als het ook nog zo is dat het nieuwe medicijn heftige bijwerkingen heeft waardoor men het alleen aan de ziekste patiënten wil geven, dan hebben de patiënten in het LUMC die het nieuwe medicijn krijgen om te beginnen al een veel lagere herstelkans. Dan is het niet vreemd dat meer mensen die het nieuwe medicijn krijgen niet herstellen, zelfs als het medicijn een positief effect heeft op de herstelkans. Was dit echter een gerandomiseerd experiment, dan kun je naar de gecombineerde data kijken en concluderen dat het nieuwe medicijn een slecht idee is.

Kookboekstatistiek

Ik begin een statistiekcollege graag met Simpsons paradox. Het vereist nog geen enkele voorkennis, maar iedereen ziet meteen dat statistiek verraderlijk kan zijn als je de context buiten beschouwing laat. Veel mensen denken dat statistiek een kookboek is met recepten die je kunt volgen om tot een uitkomst te komen, vaak een beslissing, bijvoorbeeld of we een nulhypothese wel of niet verwerpen. Zo worden statistiekvakken ook vaak gegeven bij opleidingen waar ze het moeten toepassen. Zijn de data normaal verdeeld? Volg het recept: *t*-toets. We kunnen kiezen uit verschil-

lende smaken: zijn de samples paired?, (on)gelijke variantie?, et cetera. En niet zelden zie ik op social media beslisbomen voorbijkomen met een breed scala aan hypothesetoetsen. Wat rest is op een knopje drukken in SPSS, en voilà, een wel of niet significant resultaat. Naast gevaarlijk, maakt het (toegepaste) statistiekvakken ook verschrikkelijk saai, zoals Mark van der Wiel ook laatst in een blog betoogde [2]. Ik denk dat het beter is om de principes van de statistische methodologie goed uit te leggen (bijvoorbeeld: hypothesetoetsen), met een aantal veelvoorkomende specifieke methodes (bijvoorbeeld: de *t*-toets) als voorbeelden. Om bij het hypothesetoetsen — mijn favoriete onderwerp — te blijven: wanneer je het hoe en wat snapt van het ontwerpen van methodes met als vereiste het begrenzen van een Type I fout, snap je ook sneller dat er dingen fout gaan als je niet aan de regels houdt. Zo geeft 55% van de psychologen toe [1] aan *optional stopping* [3] te doen met *p*-waarden, wat de kans op fout-positieven (gigantisch) opblaast.

Dit alles is precies wat ik zo leuk vind aan statistiek: er is altijd een interactie tussen de wiskundige theorie, en de context van de praktijk waarvoor je die theorie bedenkt. Net zoals ik denk dat het voor mensen die statistiek toepassen goed is als ze de principes van de achterliggende theorie begrijpen, denk ik dat het voor wiskundige statistici goed is, om zich te laten inspireren door de praktijk. Niet zelden geven praktische problemen aanleiding tot de meest elegante wiskundige statistiek. Ik hoop daar in volgende columns voorbeelden van te laten zien. ↩



Referenties

1 L. K. John, G. Loewenstein en D. Prelec, Measuring the prevalence of questionable research practices with incentives for truth telling, *Psychological Science* 23(5) (2012), 524–532.

2 Mark van de Wiel, <https://blog.vvsor.nl/2024/03/statistics-is-boring>.

3 Kijken naar de data of (tussen)resultaten en besluiten om wel of niet nog wat data aan het experiment toe te voegen.