Inhoudsopgave

Voorwoord

Met Begrip voor data science hebben Jan Willem en Michel op indrukwekkende wijze hun kennis en ervaring gebundeld op het gebied van het effectief en verantwoord gebruik van data in de praktijk.

Het boek is heel prettig geschreven. De toon is luchtig zonder oppervlakkig te worden, en beschrijft in consistent en verzorgd taalgebruik materie die niet iedereen even eenvoudig vindt. Dat maakt de toepassing van data science toegankelijk, en helpt ook bij een beter begrip van de technieken onder niet-cijfermatig aangelegde personen. De inhoud wordt gelardeerd met prachtige voorbeelden over de macht en zeker ook de onmacht van de moderne data sciencetechnieken. Zo passeren aanbevelingen voor cd's van Corry Konings de revue, maar ook de productiviteit van Nicolas Cage in relatie tot verdrinkingscijfers.

Waar het boek in uitblinkt is dat auteurs een heel mooie balans hebben weten te vinden in de bespreking van de theorie en de intuïtie achter data sciencemethoden, en het verantwoorde gebruik ervan in de praktijk.

Het is ondoenlijk om de theoretische achtergrond van honderden jaren ontwikkeling in data-analyse uitgebreid te bespreken in een enkel boek, en dat doen de heren dan ook niet. Er zijn heldere keuzes gemaakt in de te bespreken onderwerpen. Dat begint bij de basis: het meetniveau van variabelen. De lezer wordt ook meegenomen naar 'state-of-the-art' technieken, bijvoorbeeld rondom machinelearning. Daar waar mogelijk wordt de achtergrond van de technieken grondig uitgelegd. Op andere momenten is de keuze gemaakt om te volstaan met de intuïtieve kant. Dat draagt zeker bij aan de leesbaarheid van het boek, en ook aan de doelstelling die Jan Willem en Michel zichzelf gesteld hebben: ze willen niet dat de lezer een aap wordt die op knoppen drukt, maar hij of zij hoeft ook geen professor in de statistiek te worden!

Bij elk onderwerp is er veel aandacht voor verantwoord gebruik. Zo worden steevast de valkuilen bij de toepassing van de technieken besproken, om te voorkomen dat de lezer zich onbedoeld schuldig maakt aan lying with statistics. Maar er wordt ook uitgebreid ingegaan op de verantwoordelijkheden die het

gebruik van data met zich meebrengen op het terrein van privacy en het voorkomen van ingebakken biases. Het grote advies dat in het hele boek doorklinkt, is: blijf waakzaam!

Tegelijkertijd laten de auteurs enthousiast zien op welke manier data science echt het verschil kan maken in de praktijk. Bij de bespreking van de technieken wordt uiteengezet hoe deze het effectiefst ingezet kunnen worden en er zijn enkele cases uitgewerkt die dit op aansprekende wijze illustreren: van de inbraakmonitor van Interpolis tot aan de druktevoorspellingen van Toverland. Zo laten de auteurs zien dat het niet gaat om enkel de techniek zelf of de uitkomsten ervan, maar ook om de manier waarop deze worden gedeeld met belangrijke stakeholders.

Al met al biedt dit boek een genuanceerd en waardevol overzicht van de belangrijkste data sciencetechnieken en staat het bol van de tips & tricks waar een nieuwsgierige, zorgvuldige, open-minded en op verbetering gerichte dataprofessional zijn of haar voordeel mee kan doen. Een vlassig snorretje blijkt niet eens per se noodzakelijk.

Jan Willem en Michel eindigen met een call for action: de tijd om over implicaties van AI te spreken is nú aangebroken! Het is een thema dat momenteel al veel aandacht krijgt en in de toekomst alleen maar belangrijker zal worden. Ook al wordt gelukkig niet alles bepaald door AI: soms is het gewoon de liefde die de keuze stuurt met wie je je leven wilt delen. En daar sluit ik me van harte bij aan.

Prof. dr. Jaap E. Wieringa, Hoogleraar Onderzoeksmethoden in de Bedrijfskunde aan de Rijksuniversiteit Groningen, Onderzoeksdirecteur van het RUG Customer Insights Center

rijksuniversiteit groningen

faculteit economie en bedrijfskunde

customer insights center

“With too little data, you won’t be able to make any conclusions that you trust. With loads of data you will find relationships that aren’t real.
Big data isn’t about bits, it’s about talent.”

Douglas Merrill, voormalig CIO/VP of Engineering bij Google

Hoofdstuk 1

Data science in ons dagelijks leven

Een dag in het leven

Het is 2 minuten over 7 in de ochtend. Omdat je je wekker vaak om 2 minuten over 7 laat gaan, heeft je iPhone je gisteren de suggestie gedaan om hem ook vandaag weer op dit tijdstip af te laten gaan. Je hebt dit voorstel geaccepteerd en wordt nu dus gewekt. Je drukt je wekker uit, opent de Spotify app en kiest een afspeellijst die 'Jouw Daily Mix' heet. Hierin hoor je artiesten die je niet kent, maar waarvan de muziek je wel meteen aanspreekt, omdat deze lijkt op de muziek van artiesten waar je vaak naar luistert. Goed gemutst door deze nieuwe ontdekkingen kleed je je aan en je loopt de trap af naar beneden.

Terwijl je ontbijt, scrol je door je Facebook-tijdlijn. Hier kom je vooral berichten tegen van vrienden waar je vaak op reageert. En ook de advertenties komen je bekend voor; je komt zelfs een paar schoenen tegen dat je gisteren op de website van fabrikant Quick hebt bekeken. Je opent X, voorheen Twitter. Ook hier zie je vooral berichten van mensen van wie of organisaties waarvan je het vaakst aangeeft dat je hun berichten leuk vindt of die je het meest retweet. Klik je door naar je Apple News-feed, dan is deze gevuld met onderwerpen die aansluiten bij zaken waar je graag over leest.

Je hebt weleens gehoord dat je, door de manier waarop op sociale media berichten worden geselecteerd, in een 'bubbel' terecht kunt komen. Als je bijvoorbeeld veel op milieugerelateerde artikelen klikt, krijg je steeds meer van dit soort berichten voorgeschoteld. Klik je vaak op nieuws over de aandelenbeurs, dan krijg je veel suggesties voor artikelen die dat onderwerp belichten. Omdat je weet dat deze bubbels bestaan, heb je ook een abonnement op Blendle. Blendle doet suggesties voor artikelen die in jouw interessegebied liggen, maar ook voor onderwerpen die juist geen dagelijkse kost voor je zijn. Zo word je actief uit je bubbel gelokt, zodat je uitgedaagd wordt om uit je eigen denkpatronen en meningen te breken.

Weer helemaal op de hoogte van het nieuws ga je op weg naar je werk. Je zet Google Maps aan om afhankelijk van de drukte op de weg continu de snelste route te kunnen kiezen. Hierdoor kom je mooi op tijd aan op je werk en heb je nog tijd om even koffie te drinken met je nieuwe collega. Deze collega is een uit Syrië afkomstige vluchteling en heeft zich op voorspraak van het Centraal Orgaan Opvang Asielzoekers (COA) in de buurt van jouw kantoor gevestigd, omdat daar de kans op werk voor hem het grootst was. En dat blijkt, want hij werkt nu bij jou.

Na het koffiedrinken neem je de lift naar je werkplek. Je hoeft over het algemeen minder lang op de lift te wachten sinds je de verdieping al moet ingeven wanneer je de lift oproept, in plaats van nadat je de lift in bent gestapt. Daardoor ben je nog iets sneller op de juiste verdieping, waar je aan je bureau gaat zitten. Dit bureau is op basis van een automatisch gegenereerde plattegrond – samen met de andere bureaus – optimaal in de ruimte geplaatst om voldoende loopruimte én persoonlijke ruimte voor de werknemers te creëren.

Na een prettige dag werken ga je nog even langs de supermarkt om een paar boodschappen te doen. Om zo snel mogelijk langs de kassa te kunnen, neem je een zelfscanner uit het rek. Terwijl je door de winkel loopt, kom je een oude bekende tegen en je maakt even een praatje. Daardoor ben je lang in de winkel, terwijl je als je bij de kassa komt uiteindelijk maar vijf artikelen in je mandje hebt liggen. Door deze

Hoofdstuk 5

Makkelijk A/B-testen: de t-toets

Wanneer we het hebben over gemiddelde, modus, mediaan, variantie en standaarddeviatie, dan bevinden we ons in het rijk van de **beschrijvende statistiek**. Willen we gaan bepalen of verschillen tussen groepen statistisch significant zijn, dan wenden we ons tot de **inferentiële statistiek**.

De vraag die we met inferentiële statistiek doorgaans willen beantwoorden, is of verschillen die we tussen groepen hebben aangetroffen, door toeval tot stand kunnen zijn gekomen, of dat we met statistisch significante verschillen te maken hebben. Met 'statistisch significant' bedoelen we in essentie dat het zeer onwaarschijnlijk is dat het gevonden verschil toevallig is ontstaan.

Een van de toetsen die onder inferentiële statistiek vallen is de t-toets voor onafhankelijke steekproeven. Daar gaan we het in dit hoofdstuk over hebben.

De t-toets voor onafhankelijke steekproeven

Wellicht heb je er in een statistiekcursus in je vooropleiding al mee te maken gehad: de t-toets. Een zeer bekende, basale toets om te bekijken of de resultaten van twee groepen van elkaar verschillen. Als je het schema met de verschillende toetsen aan het eind van hoofdstuk 3 erbij pakt (figuur 3.3), dan kun je zien wanneer je deze toets gebruikt.

Je gebruikt de t-toets voor onafhankelijke steekproeven **niet** als je meerdere keren bij dezelfde groep meet; de groepen kennen ook **geen** overlap. En je gebruikt de t-toets ook **niet** als je meer dan twee groepen wil vergelijken. Daarnaast is de t-toets alleen bruikbaar als je één afhankelijke variabele wil toetsen. Oftewel: je gebruikt de t-toets als je één keer bij twee verschillende groepen één afhankelijke variabele hebt gemeten.

Voor de beeldvorming – want dit klinkt misschien nog wat abstract – geven we een praktische toepassing van de t-toets als voorbeeld. Stel dat je twee versies van een pagina op je website hebt en je wil kijken of deze versies verschil maken in het aantal producten dat je verkoopt. Je laat klanten die op je site komen aselect een van beide versies zien. Zo verdeel je je klanten in twee groepen. Na een tijdje wil je kijken of een van beide pagina's meer verkoop realiseert. Dit type experiment noemen we ook wel een A/B-test.

Je hebt bij een A/B-test inderdaad te maken met een situatie waarin je de t-toets voor onafhankelijke steekproeven kunt gebruiken. Er is tenslotte één keer gemeten (over een langere periode) bij twee verschillende, niet-overlappende groepen (klanten die de ene versie van de website hebben gezien en klanten die de andere versie hebben gezien), en je meet één afhankelijke variabele: de gerealiseerde verkoop.

Een ander voorbeeld is de gemiddelde schaderatio op de autoverzekering van mannen versus die van vrouwen. Je hebt dan twee verschillende groepen (mannen versus vrouwen) zonder overlap (niemand zit in beide groepen), waarvan je één afhankelijke variabele wil toetsen (schaderatio op de autoverzekering).

Nog een ander voorbeeld is het aantal rookmelders dat gezinnen in de stad in huis hebben versus gezinnen die op het platteland wonen. Ook hier heb je twee verschillende groepen zonder overlap met één afhankelijke variabele.

Zo kun je zelf vast ook wel voorbeelden verzinnen waarin de t-toets voor onafhankelijke steekproeven gebruikt kan worden.

Slimmer dan een aap? Lees dan verder

Je kunt je afvragen waarom je zou willen weten hoe een t-toets precies werkt. Je kunt tenslotte in Excel vrij eenvoudig een t-toets uitvoeren met bijvoorbeeld de Analysis Toolkit. Dan leer je in essentie echter niet veel meer dan als een aap op de juiste knoppen drukken. Terwijl het in de statistiek juist zo mooi en waardevol is als je doorgrondt wat er in een toets gebeurt, zodat de bruikbaarheid ervan als een puzzel voor je in elkaar valt.

Als je weet welke toetsen wanneer relevant zijn en je dit kunt koppelen aan hoe ze werken, leer je situaties herkennen waarin toetsen je kunnen helpen om de juiste inzichten te verzamelen om je besluiten op te baseren.

Het berekenen van de t-waarde

In de t-toets voor onafhankelijke steekproeven berekenen we een t-waarde. Met deze t-waarde kunnen we een oordeel vellen of een verschil dat we hebben gevonden significant is. Hoe hoger de t-waarde, des te meer reden om te veronderstellen dat het gevonden verschil significant is.

Je berekent de t-waarde en gaat die vervolgens vergelijken met een standaardtabel met zogenoemde kritische t-waardes. Deze tabel vind je in bijlage II. Is je t-waarde bij jouw gewenste significantieniveau en het aantal vrijheidsgraden dat je hebt groter dan de kritische t-waarde? Dan is je verschil significant.

Je moet dus om te kunnen vergelijken met de kritische t-waarde eerst je eigen t-waarde kunnen berekenen. Hiervoor bereken je eerst het verschil tussen de twee groeps-gemiddelden (dus bijvoorbeeld het verschil tussen de schaderatio van mannen en van vrouwen). Dit verschil deel je door de gemiddelde variabiliteit tussen de twee groepen.

De formule om de t-waarde te berekenen vind je op de volgende pagina. Deze lijkt misschien ingewikkeld, maar als je onder de breuklijn kijkt, dan zie je dat dat vooral een vermenigvuldiging is van de twee variantieformules die we in hoofdstuk 3 al hebben gezien.

Take-outs

Hoofdstuk 5

- Je gebruikt de t-toets voor onafhankelijke steekproeven als je bij twee verschillende groepen één afhankelijke variabele meet en wil weten of de groepen significant van elkaar verschillen.
- Om de t-waarde te berekenen, heb je de steekproefomvang, de gemiddeldes, de kwadraten van de verschillende waardes en de optelling van de gevonden waardes per groep nodig.
- Met de t-waarde, het significantieniveau, het aantal vrijheidsgraden en de keuze of je een- of tweezijdig toetst, kun je bepalen of verschillen significant zijn.

Alles wat je moet berekenen in een **t-toets**

t-waarde hoger dan kritische t-waarde:

H0 verwerpen

t-waarde lager dan kritische t-waarde:

H0 niet verwerpen

t-waarde

$$t = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{\sum X1^2 - \frac{(\sum X1)^2}{N1} + \sum X2^2 - \frac{(\sum X2)^2}{N2}}{N1 + N2 - 2} \left(\frac{1}{N1} + \frac{1}{N2}\right)}}$$

Kritische t-waardes

Opzoeken in tabel. Zoek deze op op basis van het aantal vrijheidsgraden.

Zoek de kritische t-waarde op in de kolom die het gewenste significatieniveau weergeeft. Let goed op of je een- of tweezijdig wil toetsen.

Bouwstenen om de t-waarde te berekenen

- $N1$ = *aantal waarnemingen in groep 1*
- $\bar{X}1$ = *optelling van alle individuele waarnemingen in groep 1, gedeeld door N1*
- $\sum X1^2$ = *optelling van alle indivduele gekwadrateerde waarnemingen in groep 1*
- $(\sum X1)^2$ = *het kwadraat van de optelling van alle individuele waarnemingen in groep 1*
- $N2$ = *aantal waarnemingen in groep 2*
- $\bar{X}2$ = *optelling van alle individuele waarnemingen in groep 2, gedeeld door N2*
- $\sum X2^2$ = *optelling van alle individuele gekwadrateerde waarnemingen in groep 2*
- $(\sum X2)^2$ = *het kwadraat van de optelling van alle individuele waarnemingen in groep 2*

Vrijheidsgraden

Vrijheidsgraden =

$$N1 + N2 - 2$$