Relevante informatie uit (sport)wetenschappelijke artikelen halen is meestal lastig. Het trekken van conclusies voor de (sport)praktijk aan de hand van deze informatie is vaak nog een stuk moeilijker. Het probleem zit vaak in de gebruikte statistiekmethode. In dit artikel wordt een aantal tekortkomingen van de traditionele methode beschreven en wordt een alternatieve methode toegelicht.

Statistisch significant of praktisch relevant?
Een andere kijk op statistiek in de (sport)wetenschap

Bas Van Hooren & Albert Smit

In de meeste wetenschappelijke artikelen staan uitspraken die de werkelijkheid niet goed beschrijven, zoals 'het innemen van bietensap zorgt voor een verbeterde sportprestatie' (statistisch significant) of 'het innemen van bietensap zorgt *niet* voor een verbeterde sportprestatie' (statistisch niet-significant). Een middenweg, zoals een onduidelijk of een neutraal effect, wordt in een wetenschappelijk artikel zelden gerapporteerd.

Het is niet zo dat wetenschappers per definitie zwart-wit denken, maar ze zijn nu eenmaal opgeleid om de resultaten op deze manier weer te geven. Daarnaast zien de *reviewers* en *editors* (de wetenschappers die de artikelen voor publicatie beoordelen) de resultaten graag gepresenteerd zoals zij het ook aangeleerd hebben gekregen. Over de werkelijke betekenis van 'statistisch significant' lijken ze liever niet te willen nadenken.

Gelukkig zijn er ook mensen die hier wel over nadenken. Hun manier van denken en rapporteren begint zich langzaam te verspreiden in de sportwetenschappelijke wereld en literatuur. Deze 'nieuwe' manier van statistiek bedrijven is echter nog geen gemeengoed. Tot het zover is zal er voornamelijk literatuur gelezen moeten worden die gebruik maakt van traditionele statistische methodieken waarbij een nulhypothese getoetst wordt.

In dit artikel zal besproken worden wat statistische significantie is, waar op gelet moet worden bij de traditionele statistiekmethode, wat de 'nieuwe' statistiekmethode inhoud en hoe deze 'nieuwe' methode gebruikt kan worden bij het interpreteren van resultaten uit de (sport)wetenschappelijke literatuur en de (sport)praktijk.

Nulhypothese significantie testen

Bij de wijdverspreide nulhypothese test gebruikt de onderzoeker statistiekssoftware om een p-waarde ('probability') te produceren. Deze p-waarde is de waarschijnlijkheid van het verkrijgen van elke waarde groter dan het geobserveerde effect (onafhankelijk of dat positief of negatief is), als de nulhypothese waar zou zijn. Bij een p-waarde kleiner dan 0.05 ($p < .05$) of een 95% betrouwbaarheidsinterval dat niet de waarde nul bevat, wordt de nulhypothese verworpen en wordt van de uitkomst gezegd dat deze statistisch significant is. Hierbij wordt er vanuit

gegaan dat er een kleine kans is dat het gevonden effect op toeval berust. Begrijpt u het nog? In een poging betekenis te geven aan deze mysterieuze aanpak, interpreteren onderzoekers de p-waarde vaak foutief, namelijk als de waarschijnlijkheid dat de nulhypothese waar is. Het gevolg is dat ze hun resultaten ook fout interpreteren.
Neem het volgende voorbeeld: er wordt een onderzoek uitgevoerd naar het effect van krachttraining op spronghoogte. De nulhypothese hierbij is: vier weken krachttraining zorgt voor géén statistisch significante verandering in de spronghoogte. Door middel van statistische testen wordt nagegaan of de nulhypothese verworpen mag worden. Bij $p < .05$ of een 95% betrouwbaarheidsinterval dat niet de waarde 0 bevat wordt geconcludeerd dat er een statistisch significant verschil is gevonden en wordt de nulhypothese verworpen. Hierbij wordt ook aangenomen dat er een kleine kans is dat de verandering in spronghoogte door invloed van krachttraining op toeval berust. Vervolgens wordt aangenomen dat de alternatieve hypothese waar is: vier weken krachttraining zorgt wél voor een statistisch significante verandering in de spronghoogte. Deze conclusie mag echter niet getrokken worden op basis van deze gegevens[1], aangezien concluderen dat het ene niet waar is niet automatisch betekent dat het tegenovergestelde dan wel waar moet zijn. Daarnaast weet je simpelweg nog niet of het statistisch gevonden verschil verklaard kan worden door die krachttraining. Dat is hetzelfde als onderzoek doen naar het dopinggebruik onder wielrenners met als nulhypothese 'geen enkele wielrenner gebruikt doping'. Als deze nulhypothese verworpen wordt op basis van het statistisch onderzoek zou de conclusie 'alle wielrenners gebruiken doping' luiden. Dat is immers het tegenovergestelde van de nulhypothese. Echter, op basis van het statistisch onderzoek zou de conclusie zou moeten zijn: 'Er is niet aangetoond dat geen enkele wielrenner doping gebruikt'. Eigenlijk is het antwoord op de vraag 'hoeveel wielrenners gebruiken er nu eigenlijk wèl doping?' ook veel interessanter. Hetzelfde geldt voor het onderzoek naar het effect van krachttraining op spronghoogte. De enige conclusie die daar getrokken mag worden is dat er *niet* aangetoond kan worden dat spronghoogte door krachttraining *niet* verandert (maar er gebeurt ALTIJD iets, ook als er bij wijze van spreken vier weken pim-pam-pet gespeeld zou worden). Ook hier is het eigenlijk interessanter om te weten hoeveel hoger er wordt gesprongen als gevolg van de krachttraining. Hier komen we later nog op terug.

Significant versus (praktisch) relevant

Men kan zich ook afvragen wat er zo speciaal is aan het getal $p < .05$. Het antwoord hierop is: eigenlijk niets. De invloedrijke statisticus Fisher heeft ooit eens gezegd: 'We shall not often go astray if we draw a conventional line at 0.05', maar daar bedoelde hij niet mee dat het als een absolute regel gezien moet worden.[2] Dat deze waarde een soort van heilige grens geworden is, was zeker niet zijn bedoeling: '… surely, God loves the .06 nearly as much as .05'.[3]
De meest gehoorde reden om p- waarden te gebruiken is om te toetsen of het gevonden verschil op toeval berust. Zo wordt vaak aangenomen dat er bij $p < .05$ een waarschijnlijkheid is van maximaal 5% dat het gevonden resultaat op toeval berust. Deze aanname is echter niet terecht.[4-6] Een andere onterechte aanname is dat er bij $p > .05$ of een 95% betrouwbaarheidsinterval met de waarde 0 aangetoond is dat de interventie geen of nauwelijks effect heeft. Een gebrek aan bewijs voor een effect is echter niet hetzelfde als bewijs voor geen effect, aldus Altman en Bland[7]: 'Absence of evidence is not evidence of absence'.
Een combinatie van een betrouwbare meting en een grote steekproef (veel deelnemers in het onderzoek) zal vaak leiden tot statistisch significante resultaten. Deze resultaten zijn echter lang niet altijd praktisch relevant. Andersom zijn resultaten die statistisch niet significant zijn mogelijk wel praktisch of klinisch relevant. Statistische significantie en praktische of klinische relevantie zijn dus verschillende concepten, die helaas vaak door elkaar gehaald worden.
Neem een voorbeeld uit het wielrennen. Bij goed getrainde wielrenners en triatleten werd na het toedienen van een combinatie van glucose en aminozuren een niet significante verbetering van 2,9 minuten gevonden op een 160 minuten durende tijdrit. Er werd geconcludeerd dat het toedienen van deze supplementen geen effect had.[8] Het is echter goed mogelijk dat een verbetering van deze grootte (1.8%) een relevant effect heeft op de prestatie van goed getrainde atleten. Zo werd het werelduurrecord recent net niet verbroken door Thomas Dekker. Hij kwam 270 meter tekort voor het record en met een gemiddelde snelheid van zo'n 52 km per uur komt dit neer op 18.7 seconden. 18.7 seconden op 60 minuten is slechts 0,5%. Een interventie waarmee 1.8% sneller gereden had kunnen worden was daarom waarschijnlijk praktisch relevant geweest.
Naast de zojuist besproken verkeerde interpretaties zijn er nog verschillende voorbeelden van verkeerde interpretaties van p-waarden en tekortkomingen van het nulhypothese testen.[4,5] De vele tekortkomingen van de nulhypothese hebben al sinds de invoer door Neyman en Pearson tot veel kritiek (o.a. [9-13]) geleid. Ondanks deze kritiek wordt de methode nog steeds veel toegepast en zelfs aangeleerd aan studenten in cursussen en op universiteiten. Dit komt onder andere doordat

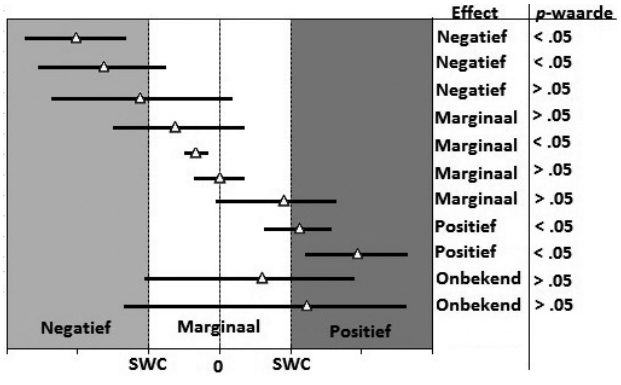
Absence of evidence is not evidence of absence

Er bestaat in het wetenschappelijk onderzoek iets dat 'publication bias' genoemd wordt. Het komt er op neer dat niet alle onderzoeken gepubliceerd worden. Vooral onderzoeken die ofwel een statistisch significant effect vinden, ofwel een bepaald experiment voor de eerste keer uitvoeren (en dus geen herhaling van een eerder onderzoek zijn) worden gepubliceerd. Onderzoeken waarin geen statistisch significant effect is gevonden worden nauwelijks gepubliceerd. Daar komt bij dat het eerste artikel over een nieuw onderwerp met een positief effect de grootste invloed heeft, ondanks dat er daarna meerdere vergelijkbare (maar nooit precies dezelfde) onderzoeken worden gepubliceerd die dit effect niet vinden. Denk hierbij aan het onderzoek naar de werking van bietensap.
Verder wordt er maar weinig onderzoek met topsporters gedaan, omdat er geen geld voor is en omdat er simpelweg weinig topsporters zijn die willen deelnemen aan een onderzoek. Dit maakt conclusies trekken uit wetenschappelijk onderzoek nog lastiger, zeker als het om topsporters gaat. Als er geen bewijs is voor de werking van 'iets' bij een (top) sporter wil dit daarom niet zeggen dat dit bij hen per definitie niet werkt. Dit geldt natuurlijk ook andersom. Met MBI kan in ieder geval gekeken worden hoe groot het gevonden effect is en kan worden ingeschat hoe reëel het is dat de interventie ook een relevant effect heeft op een beter getrainde groep.

veel mensen niet bekend zijn met alternatieve (betere) methodes. Een van die methodes is 'magnitude-based inferences' (MBI).[14,15] Hoewel ook deze methode geen perfecte oplossing biedt voor alle statistische problemen is een aantal belangrijke aspecten verbeterd ten opzichte van het nulhypothese testen. Daardoor is het zowel voor de (sport)praktijk als de (sport)wetenschap een zeer geschikte methode.

Magnitude-based inferences (MBI)

De statistische methode MBI is direct gebaseerd op de onzekerheid van de echte waarde in de statistiek en daarmee intuïtiever en praktischer dan de nulhypothese test. Een korte samenvatting van MBI:
Er wordt vanuit gegaan dat elke meting een momentopname is en dat de werkelijke waarde met een bepaalde onzekerheid ergens in een bepaald gebied (confidence limits/intervals) rondom de gemeten waarde ligt.
Deze onzekerheid wordt vervolgens gebruikt voor het bepalen van de praktische of klinische relevantie, door te berekenen wat de kans is dat de waarde substantieel positief of negatief is. Als de waarschijnlijkheid van de werkelijke waarde zowel in het positieve als het negatieve deel ligt, dan is de uitkomst onduidelijk. Anders is het effect substantieel positief, verwaarloosbaar of negatief (figuur 1).
Een voorbeeld: na vier weken sprinttraining worden veertig voetballers getest op de twintig meter sprintprestatie (momentopname). De spelers zijn gemiddeld 0.16 seconden sneller geworden. Twee spelers zijn echter 0.02 seconde langzamer geworden. Uit de statistische analyse met MBI blijkt dat 1) de kans op een achteruitgang klein is en 2) als er een achteruitgang optreedt deze ook nog eens verwaarloosbaar klein zal zijn. Het trainingsprogramma lijkt daarmee een substantieel positief effect te hebben.



Figuur 1 Negatieve, marginale en positieve effecten aan de hand van betrouwbaarheidsintervallen en SWC.[14]

Smallest worthwhile change (SWC)

Om over effecten uitspraken te kunnen doen als 'substantieel positief', 'verwaarloosbaar' of 'negatief' moet vooraf bepaald zijn welke kleinste verandering nog de moeite waard is. Dit is de zogeheten 'smallest worthwhile change', afgekort SWC. Het is de kleinste verandering die in de praktijk ook werkelijk invloed heeft op de resultaten, bijvoorbeeld het opschuiven van een plek in het klassement. In de topsport is de SWC voor veel parameters die het resultaat direct beïnvloeden ongeveer 1%.[16]
Voor sporten waarin op tijd of afstand afgerekend wordt, wordt de SWC bepaald door het berekenen van de typische variatie in eindtijd of afstand over een aantal wedstrijden, rekening houdend met omstandigheden, et cetera. Bij veel variatie in de prestatie is de SWC groter en bij weinig variatie kleiner. Bij veel variatie is er immers een grotere verandering nodig om deze te overtreffen.
Een voorbeeld: als de eindtijd bij 200m sprintwedstrijden iedere keer met één seconde varieert (een groot verschil), dan moet de SWC groter zijn dan één seconde om er zeker van te zijn dat de atleet iedere keer wint. Als de eindtijd echter met slechts enkele honderdsten van een seconde verschilt, dan zal de

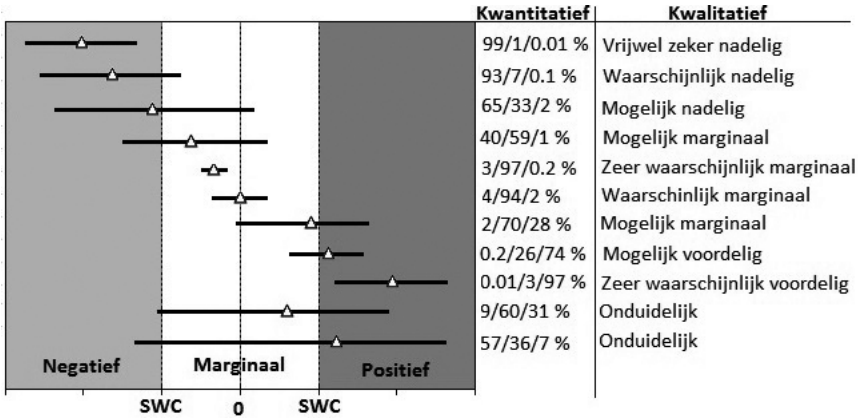
SWC ook enkele honderdsten van een seconde zijn.
De totale variatie kan worden uitgedrukt als een coëfficiënt van de variatie (CV). Bij een CV van 1% betekent dit één cm per honderd cm, 0.1 sec per tien sec, et cetera. Dit kan echter beter niet zelf berekend worden, omdat het statistisch nogal ingewikkeld is. Voor een aantal sporten is de variatie al bepaald en uitgedrukt in een CV. Bij individuele sporters wordt de SWC berekend als 0.3 keer de CV (voorheen werd 0.5 aangehouden, maar deze grens is ondertussen bijgesteld).[17] De SWC wordt op deze manier berekend omdat een verandering die groter is dan de typische variatie nagenoeg altijd een relevante prestatieverbetering oplevert. Omdat het verschil tussen een eerste, tweede of derde plek soms erg klein is, zijn kleine veranderingen ook al relevant. Door de CV met 0.3 te vermenigvuldigen worden deze veranderingen als relevant gezien.
Bij internationale topsprinters in de atletiek ligt de CV rond de 0.6%. Een verbetering van 0.3 keer deze variatie (= 0.18%) zou voor deze atleten al een relevant verschil betekenen.[18,19] Voor toptriatleten is de SWC 0.33% van de totale racetijd, of 0.24% als alleen de variatie van het hardloopdeel wordt meegerekend.[16] Voor een overzicht van CV-waarden om de SWC te berekenen wordt verwezen naar een artikel door Smit.[20]
Bij teamsporten bestaat er geen directe relatie tussen een veranderde prestatie in een bepaalde meting en de wedstrijdprestatie. Zo kan een volleyballer bijvoorbeeld 3 cm hoger zijn gaan springen als gevolg van een bepaalde trainingsvorm, maar omdat het spel ook wordt beïnvloed door zijn teamgenoten en omdat zijn eigen presteren ook afhankelijk is van zijn tactische en technische vaardigheden, is niet duidelijk of zijn toegenomen spronghoogte ook direct de winkansen van zijn team beïnvloedt. De SWC in teamsporters wordt daarom berekend aan de hand van een gestandaardiseerde 'effect size' van 0.20, ook bekend als Cohen's d (zie verdere toelichting onder het kopje 'Kwalitatief rapporteren van het effect').[17] Deze is te berekenen door de standaarddeviatie van een testwaarde van de gehele groep te vermenigvuldigen met 0.20. De waarde die dan verkregen wordt is de SWC in absolute waarden.
De SWC geeft duidelijkheid over de grens die overschreden moet worden om te spreken van een substantiële verbetering of verslechtering, of van een verwaarloosbaar effect. Door kennis te nemen van de SWC is men veel beter in staat om zinvolle uitspraken te doen over de effecten van een interventie.

Verschil tussen nulhypothese testen en MBI bij de interpretatie van resultaten

Bij MBI worden de betrouwbaarheidsintervallen (zie kader) geïnterpreteerd in relatie tot de SWC en niet - zoals bij nulhypothese testen - in relatie tot de 0-waarde. De resultaten verschillen daarom ook van elkaar, zoals te zien is in figuur 1. Het vijfde interval van boven is bijvoorbeeld statistisch significant, want het 95% betrouwbaarheidsinterval bevat niet de waarde 0. Deze verandering is echter kleiner dan de SWC, dus niet praktisch relevant.
Het zevende betrouwbaarheidsinterval van boven bevat wel de waarde 0, maar het is duidelijk dat er geen negatief, maar wel een positief effect verwacht kan worden. Afhankelijk van de kosten (geld, tijd, moeite etc.) zou men dus kunnen beslissen de interventie te gebruiken.
Door de betrouwbaarheidsintervallen te interpreteren in relatie tot de SWC is tevens te zien hoe dit leidt tot drie mogelijke opties: een negatief, marginaal of positief effect. Als een betrouwbaarheidsinterval in zijn geheel in het positieve deel ligt is duidelijk dat de interventie een positief effect heeft. Bij een betrouwbaarheidsinterval waarbij de grenzen in zowel het positieve als negatieve deel liggen is het effect van de interventie niet duidelijk. Meer onderzoek, met een grotere steekproef of betere metingen, is nodig om de grenzen van het betrouwbaarheidsinterval dichter bij elkaar te brengen. Ook bij een onduidelijk effect kan worden aangegeven welk effect het meest waarschijnlijk is (bijvoorbeeld: 'onduidelijk, maar mogelijk positief'). Zodra de grenzen van het betrouwbaarheidsinterval in twee delen liggen, is duidelijk dat de derde optie niet van toepassing is. Liggen de grenzen bijvoorbeeld in het marginale en positieve deel, dan is duidelijk dat het effect niet negatief zal zijn. Maar welke van de andere twee effecten (marginaal of positief) waar-

Betrouwbaarheidsintervallen

Als bij iemand een spronghoogte van 45.7 cm wordt gemeten, zal de echte waarde waarschijnlijk niet exact 45.7 cm zijn. Bij iedere meting is er immers sprake van een bepaalde meetfout. Om hier rekening mee te houden worden betrouwbaarheidsintervallen gebruikt. Een betrouwbaarheidsinterval is het bereik waarin de werkelijke waarde van het individu (of een groep) waarschijnlijk valt. 'Waarschijnlijk' duidt hierbij meestal op 95%. Batterham en Hopkins[14] raden echter aan om bij het testen van topsporters 90% betrouwbaarheidsintervallen aan te houden.
Een voorbeeld: er worden twee groepen vergeleken op spronghoogte. Tussen de groepen is een verschil van 12 cm, met een 90% betrouwbaarheidsinterval dat loopt van 8 tot 16 cm. Het werkelijke verschil tussen de groepen ligt dan met 90% zekerheid ergens tussen 8 en 16 cm.



Figuur 2 Waarschijnlijkheid dat het gevonden effect gelijk is aan het echte effect.[14]

schijnlijker is? Om hier duidelijkheid over te geven wordt aangegeven hoe waarschijnlijk het is dat het gevonden effect dezelfde waarde heeft als het echte effect (zie figuur 2).
Bij onderzoeken naar slechts enkele parameters kunnen de exacte percentages gerapporteerd worden (zie de linker tekstkolom in figuur 2 & tabel 1), maar bij onderzoeken naar meerdere parameters gaat de voorkeur uit naar kwalitatieve beschrijvingen (rechter tekstkolom figuur 2, tabel 1). Deze manier van rapporteren geeft snel duidelijkheid over het gevonden effect. Bovendien is de kwalitatieve beschrijving gemakkelijk te interpreteren. De percentages kunnen met kant-en-klare spreadsheets (zie verderop) automatisch worden omgezet in een kwalitatieve beschrijving.

kwantitatieve verandering	kwalitatieve beschrijving
< 0.5%	hoogst onwaarschijnlijk, vrijwel zeker niet
0.5% - 5%	zeer onwaarschijnlijk
5 - 25%	onwaarschijnlijk, waarschijnlijk niet
25 - 75%	mogelijk
75 - 95%	waarschijnlijk
> 99.5%	zeer waarschijnlijk

Tabel 1 Kwantitatieve veranderingen en kwalitatieve beschrijvingen.[21]

Kwalitatief rapporteren van het effect

Het laatste aspect waar rekening mee gehouden moet worden is de grootte van het effect. Ook dit kan met kwalitatieve begrippen gerapporteerd worden. Hierdoor is makkelijker af te lezen hoe een effect geïnterpreteerd moet worden. Als de gemiddelde gesprongen reikhoogte van een groep volleyballers bijvoorbeeld met 5.4 cm toeneemt (van 278.3 cm naar 283.7 cm), is dit dan een kleine, middelgrote of grote verandering? Door de effect size te berekenen (0.29) en de kwalitatieve beschrijving uit tabel 2 te gebruiken kan dit bepaald worden. De effectgrootte wordt berekend aan de hand van de effect size (Cohen's d). Dit is de gestandaardiseerde verandering in gemiddelden, een getal dat aangeeft hoeveel standaarddeviaties twee groepen of twee gemiddelden uit elkaar liggen. In het voorbeeld hierboven is het gemiddelde met 0.29 standaarddeviaties verbetert. In tabel 2 is te zien of dit dan als een marginale, kleine of grote verbetering gezien mag worden. Deze schaal is overigens gebaseerd op onderzoeken in de gedragswetenschappen. Om de effecten van bijvoorbeeld krachttraining te beschrijven kan wellicht beter een andere schaal gebruikt worden.[22]
Soms worden in wetenschappelijke artikelen percentages gebruikt om de grootte van een effect uit te drukken (bijvoorbeeld '10% verbetering in spronghoogte'). Bij het gebruik van percentages wordt echter geen rekening gehouden met de variantie tussen de individuen, terwijl dit bij het gebruik van de effect size wel wordt gedaan.

kwantitatieve verandering	kwalitatieve beschrijving
< .02	marginaal
0.2 - 0.6	klein
0.6 - 1.2	middelgroot
1.2 - 2.0	groot
2.0 - 4.0	zeer groot
> 4.0	extreem

Tabel 2 Effect sizes met kwalitatieve beschrijving.[23]

Door alle informatie te combineren kan een veel genuanceerdere uitspraak gedaan worden over de effecten van een interventie in vergelijking met nulhypothese testen. Als op basis van een nulhypothese test bijvoorbeeld wordt geconcludeerd 'dat er geen significant ($p > .05$) verschil is', dan zou de conclusie op basis van MBI kunnen luiden 'dat de interventie waarschijnlijk een marginaal negatief effect heeft en mogelijk een groot positief effect'.
Het voordeel ten opzichte van nulhypothese testen is dat er geen zwart-wit antwoord wordt gegeven op de vraag of er wel of geen effect is. Een interventie kan met een bepaalde zekerheid een bepaalde effectgrootte hebben. Op basis van deze informatie kan de coach beslissen of een interventie toegepast moet worden. Hierbij kan afgewogen worden of het mogelijke negatieve

effect opweegt tegen het mogelijke positieve effect. Ook de kosten van de interventie en de individuele respons erop kunnen in de afwegingen worden meegenomen. Zo is een goedkoop (dopingvrij) voedingssupplement met een zeer onwaarschijnlijk marginaal negatief effect, maar waarschijnlijk ook een groot positief effect een waardevolle toevoeging aan de winstkansen van een atleet. Aan de andere kant is een vorm van training die waarschijnlijk een marginaal tot klein positief effect heeft, maar per week wel zes uur extra inspanning kost, de moeite misschien niet waard.

Waarom is MBI nog geen algemeen gebruikte methode in de (sport)wetenschap?

Men kan zich afvragen waarom MBI nog niet overal gebruikt wordt als deze methode zowel voor de praktijk als de wetenschap geschikter is dan de traditionele methode van nulhypothese testen. Dit komt omdat veel mensen nog niet bekend zijn met deze methode en omdat er redelijk wat weerstand is tegen nieuwe statistische methoden.[24] Deze weerstand komt onder andere door de auteurs en redacteuren van wetenschappelijke tijdschriften. Omdat nog niet alle wetenschappelijke tijdschriften de auteurs verplichten om betere statistische methodes toe te passen zullen docenten statistiek hun lessen en tekstboeken niet snel aanpassen. Ook de nieuwe generatie studenten zal zo traditionele statistiekmethoden aangeleerd krijgen en weer weerstand bieden tegen verandering. Gelukkig zijn er steeds meer auteurs die de waarde inzien van MBI en ook steeds meer wetenschappelijke tijdschriften accepteren geen artikelen meer met nulhypothese testen.
Een andere oorzaak van de weerstand tegen MBI is dat de meeste statistiekprogramma's niet gebruikt kunnen worden, omdat deze (nog) niet alle berekeningen kunnen uitvoeren die nodig zijn bij deze methode. Daarom zijn er kant-en-klare spreadsheets ontwikkeld, die gratis gedownload kunnen worden op de website (www.sportsci.org) van professor Hopkins, de ontwikkelaar van MBI. Tevens is bij iedere spreadsheet informatie beschikbaar over hoe deze gebruikt moet worden. In tabel 3 is een kort overzicht te zien van spreadsheets die gebruikt kunnen worden voor bepaalde doeleinden. Door de tabel van links naar rechts te lezen kan men zien welke klassieke test toegepast zou worden bij een bepaalde onderzoeksopzet en welke MBI spreadsheet dit doel vervult. In de onderste rij van de tabel staat een verwijzing naar een spreadsheet waarmee een p-waarde omgezet kan worden in 'confidence limits voor inference' waarden over de werkelijke waarde van een effect. Op deze manier kan bestaand onderzoek toch in het kader van praktische relevantie bekeken worden.

Nadelen

Ook MBI is geen perfecte methode zonder tekortkomingen. In een recent verschenen review[25] wordt kritiek geleverd op de methode. Een deel van die kritiek is in een reactie[26] weerlegd. Een nadeel van MBI (en statistiek in het algemeen) is, dat er vaak gekeken wordt naar groepsgemiddelden, terwijl een individuele aanpak erg belangrijk is. Zeker in de topsport waar kleine veranderingen het verschil kunnen maken tussen winnen en verliezen. Resultaten van MBI relateren aan de echte waarde voor de *populatie*, niet aan de waarden van een *individu* in de populatie. Het is mogelijk dat een interventie vrijwel zeker een voordelig effect heeft op een grote groep, maar voor een enkel individu toch nadelig is.[14] Richtlijnen om de individuele effecten te monitoren in combinatie met MBI zijn inmiddels verschenen[27], maar het valt buiten het bestek van dit artikel om deze te bespreken. In een eerder in *Sportgericht* verschenen artikel[20] wordt wel beschreven hoe MBI in combinatie met de meetfout van de test gebruikt kan worden om individuele vooruitgang te monitoren.

Conclusie

Het traditionele nulhypothese testen heeft vele tekortkomingen, die kunnen leiden tot verkeerde conclusies. MBI heeft enkele van deze nadelen niet. Hierdoor is deze methode een stuk geschikter om te gebruiken bij het onderzoek. De resultaten van MBI worden op een praktischer manier gerapporteerd, waardoor ze veel makkelijker te interpreteren zijn. Het gat tussen wetenschap en praktijk kan zo

Tabel 3 Overzicht van onderzoeksdoelen met de bijbehorende 'klassieke' statistische testen respectievelijk spreadsheets voor de MBI methode.

doel onderzoek	klassieke test	spreadsheet MBI
vergelijken twee gemiddeldes	*t*-test	www.sportsci.org/resource/stats/xCompare2groups.xls
vergelijken meerdere gemiddeldes	ANOVA	www.sportsci.org/resource/stats/xPostOnlyCrossover.xls www.sportsci.org/resource/stats/xPrePostCrossover.xls
vergelijken gemiddeldes over langere periode	repeated-measures	www.sportsci.org/resource/stats/xParallelGroupsTrial.xls
betrouwbaarheid	Pearson/Spearman correlatie	www.sportsci.org/resource/stats/xrely.xls
validiteit	regressie	www.sportsci.org/resource/stats/xvalid.xls
omzetten *p*-waarde in confidence limits en inferences	n.v.t.	www.sportsci.org/resource/stats/xcl.xls

verkleind worden.

Besluit

Wij hebben niet de illusie dat wij de wetenschappelijke wereld kunnen veranderen. Wij hopen echter wel dat we onderzoekers en auteurs (bijvoorbeeld in *Sportgericht*) kunnen enthousiasmeren om de klinische en praktische relevantie van hun onderzoeksresultaten te rapporteren, zodat de lezer deze beter kan interpreteren en gebruiken.

Referenties

1. Wilkinson M (2014). Distinguishing between statistical significance and practical/clinical meaningfulness using statistical inference. *Sports Medicine*, 44 (3), 295-301.

2. Fisher RA (1992). Statistical methods for research workers. In: Kotz S & Johnson N (eds.), *Breakthroughs in Statistics*, pp. 66-70. New York: Springer.

3. Rosnow RL & Rosenthal R (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44 (10), 1276-1284.

4. Kline RB (2004). *What's wrong with statistical tests and where we go from here*. Washington, DC: APA books.

5. Schmidt FL & Hunter JE (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: Harlow LL, Mulaik SA & Steiger JH (eds.), *What if there were no significance tests*, pp. 37-64. Mahwah, NJ: Erlbaum.

6. Nuzzo R (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, 506, 150-152.

7. Altman DG & Bland JM (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*, 311 (7003), 485.

8. Madsen K et al. (1996). Effects of glucose, glucose plus branched-chain amino acids, or placebo on bike performance over 100 km. *Journal of Applied Physiology*, 81 (6), 2644-2650.

9. Ziliak ST & McCloskey DN (2008). *The cult of statistical significance*. Ann Arbor: University of Michigan Press.

10. Drinkwater E (2008). Applications of confidence limits and effect sizes in sport research. *Open Sports Sciences Journal*, 1, 3-4.

11. Stang A, Poole C & Kuss O (2010). The ongoing tyranny of statistical significance testing in biomedical research. European Journal of Epidemiology, 25 (4), 225-230.

12. Hopkins WG et al. (2011). Statistical perspectives: all together NOT. *The Journal of Physiology*, 589 (21), 5327-5329.

13. Cumming G (2014). The new statistics: Why and how. Psychological Science, 25 (1), 7-29.

14. Batterham AM & Hopkins WG (2006). Making meaningful inferences about magnitudes. *International Journal of Sports Physiology and Performance*, 1 (1), 50-57.

15. Hopkins WG et al. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine & Science in Sports & Exercise*, 41 (1), 3-13.

16. Paton CD & Hopkins WG (2005). Competitive performance of elite Olympic-distance triathletes: reliability and smallest worthwhile enhancement. *Sportscience*, 9, 1-5.

17. Hopkins WG (2004). How to interpret changes in an athletic performance test. *Sportcience*, 8, 1-7.

18. Hopkins WG, Hawley JA & Burke LM (1999). Researching worthwhile performance enhancements. *Sportscience*, 3 (1).

19. Hopkins WG (2005). Competitive performance of elite track-and-field athletes: variability and smallest worthwhile enhancements. *Sportscience*, 9, 17-20.

20. Smit A (2009). Betrouwbare inspanningstesten: simpele statistiek voor interpreteerbare resultaten. *Sportgericht*, 63 (1), 40-45.

21. Hopkins WG (2007). A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. *Sportscience*, 11, 16-20.

22. Rhea MR (2004). Determining the magnitude of treatment effects in strength training research through the use of the effect size. *The Journal of Strength & Conditioning Research*, 18 (4), 918-920.

23. Hopkins WG (2002). A scale of magnitudes for effect statistics. http://www.sportsci.org/resource/stats/effectmag.html (1 maart 2015).

24. Sharpe D (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, 18 (4), 572-582.

25. Welsh AH & Knight EJ (2015). "Magnitude-based inference": a statistical review. *Medicine & Science in Sports & Exercise*, 47 (4), 874-884.

26. Batterham AM & Hopkins WG (2015). The case for magnitude-based inference. *Medicine & Science in Sports & Exercise*, 47 (4), 885.

27. Hopkins WG (2015). Individual responses made easy. *Journal of Applied Physiology*, 118 (12), 1444-1446.

Over de auteurs

Bas Van Hooren is afgestuurd als bewegingsdeskundige aan Fontys Sporthogeschool Eindhoven. Momenteel volgt hij een master bewegingswetenschappen aan de Universiteit van Maastricht. Tevens is hij op freelance basis werkzaam als fysieke trainer voor topsporters en topsporttalenten in voornamelijk Zuid-Limburg. E-mail: basvanhooren@hotmail.com.

Albert Smit is bewegingswetenschapper en werkte 11 jaar lang voor NOC*NSF als embedded scientist en inspanningsfysioloog. Momenteel werkt hij als zelfstandig sportfysioloog voor meerdere opdrachtgevers. E-mail: albert.smit77@icloud.com, website: www.albertwot.nl.