Marjolein Bolten wordt gemonitord tijdens haar looptraining. Foto: Rikkert Harink

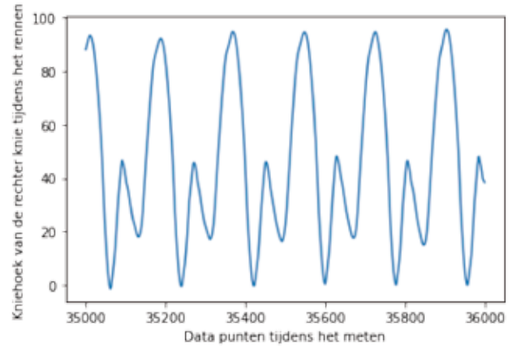
Welke test ‘loopt’ het best?

Marjolein Bolten

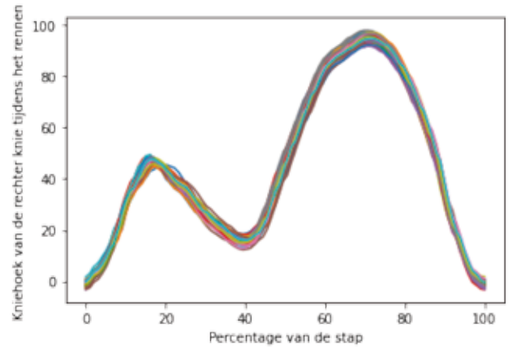
Hardlopen is een populaire en laagdrempelige sport. Een paar hardloopschoenen is al voldoende om een rondje te lopen. Het kan op jouw moment of in een gezellige groep. Bovenal, het is aantoonbaar gezond en geestverruimend. Echter, de voordelen kennen ook een nadeel, blessures. Eenderde van de blessures ontstaan tijdens het hardlopen, betreft een knieblessure. Deze ontstaan vaak door overtraining en/of te weinig rust na de training, oftewel het gaan trainen terwijl de spieren nog vermoeid zijn (Mechelen, 1992).

In het kader van deze blessures is het zinvol om looptrainingen te monitoren en bijvoorbeeld te kijken naar de hartslag of spierbewegingen. Het statistisch analyseren van de data kan helpen om de blessures te voorkomen. Elke stap is anders en daardoor zullen de bewegingspatronen van meerdere stappen niet perfect over elkaar heen lopen. Deze variatie kan gemodelleerd worden als statistische ruis, zodat de data kunnen worden beschreven als de echte gemiddelde stap plus deze statistische ruis.

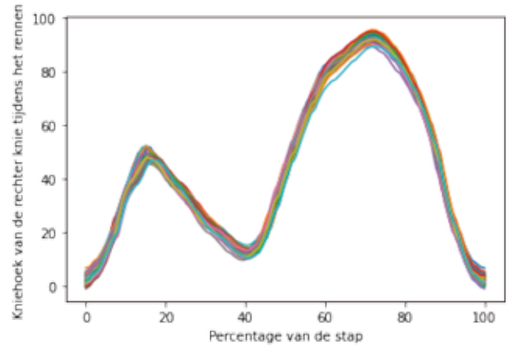
Het analyseren van spierbewegingen is een stuk moeilijker dan het analyseren van een hartslag. Dit komt doordat spierbewegingen in feite kunnen worden gezien als een tijdreeks, waarbij meerdere functies achter elkaar de data beschrijven. Waar bij de hartslag wordt gekeken naar een reeks van getallen voor het analyseren, wordt bij spierbewegingen gekeken naar een reeks van trajecten, waarbij een traject bestaat uit meerdere getallen die bij elkaar horen. Voor zulke tijdreeksen zijn er minder analyse-



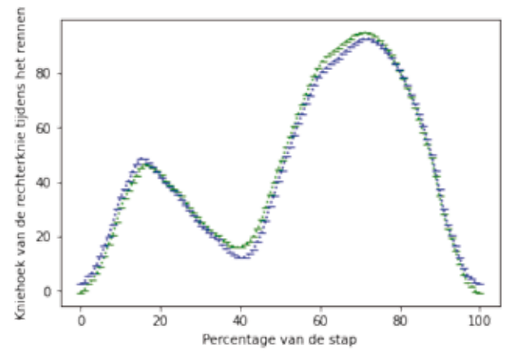
Figuur 1. Ruwe data



Figuur 2. Niet vermoeid



Figuur 3. Vermoeid



Figuur 4. Gemiddelde trajecten

mogelijkheden bekend dan voor datasets met reeksen van getallen.

Figuur 1 toont de bewegingen van de knie in het sagittale vlak, het anatomische vlak dat het lichaam verdeeld in een links en rechts vlak, tijdens het hardlopen. Deze bewegingen kunnen worden gezien als een tijdreeks. In dit artikel worden drie verschillende statistische methodes vergeleken en uitgevoerd op de data om te kijken welke methode het meest geschikt is voor het analyseren van de bewegingen van de kniehoeken tijdens het hardlopen.

Data

De data zijn verzameld tijdens een vermoeidheidsrun op een loopband waarbij de hardloper zo lang mogelijk moest hardlopen op 103% van zijn 8 km tempo. Door op dit tempo te lopen was vermoeidheid gegarandeerd aan het eind van het protocol. De eerste 60 stappen worden gebruikt als data groep 1, de niet-vermoeide groep en de laatste 60 stappen worden gebruikt als data groep 2, de vermoeide groep. De data zijn weergegeven in figuren 2 en 3, met groep 1 in figuur 2 en groep 2 in figuur 3, waarbij elke 60 stappen over elkaar heen geplot zijn om het verschil te verduidelijken.

Zoals in figuur 2 en 3 te zien is, zijn er verschillen tussen de groepen, maar de vraag is of dit verschil te verklaren is door de aanwezigheid van ruis of dat er een significant verschil is. Om deze vraag te beantwoorden zijn er drie verschillende statistische methodes geanalyseerd door te testen op verschil in de gemiddelde trajecten van de twee groepen. Hiervoor zijn dus eerst de gemiddelde trajecten berekend en deze zijn te zien in figuur 4.

In de statistische methodes wordt gebruik gemaakt van de volgende nul- en alternatieve hypothese:

H_0: De twee gemiddelde trajecten zijn identiek

H_1: De twee gemiddelde trajecten zijn niet identiek

De statistische methodes zullen met een 95% betrouwbaarheidslevel de nulhypothese accepteren of verwerpen. Door het verwerpen van de nulhypothese kan geconcludeerd worden dat de twee trajecten niet identiek zijn.

Zoals in figuur 4 te zien is, zijn deze twee gemiddelde trajecten niet identiek aan elkaar, maar welke methodes zijn in staat dit verschil ook te detecteren en zullen dus de nulhypothese verwerpen?

Methoden

De drie statistische methoden die vergeleken worden in dit artikel zijn, door middel van A. betrouwbaarheidsintervallen, B. Tweezijdige t-test en C. Bootstrap test. De eerste twee methoden vallen onder de categorie gecombineerde testen. Dat betekent dat er meerdere testen tegelijkertijd naast elkaar uitgevoerd worden met uiteindelijk één conclusie: de nulhypothese verwerpen of accepteren. Dit is een simpele combinatie van univariate methoden, die niet de specifieke structuur van het hele traject meenemen. De laatste methode is geen gecombineerde test en test wel direct op het hele traject.

Voor de eerste twee testen is elke stap in 100 individuele punten opgedeeld, waarbij op elk van die losse punten de betreffende methode is uitgevoerd. Zo zijn er 100 betrouwbaarheidsintervallen opgesteld die, na een multipliciteit correctie (zie volgende paragraaf), samen de 95%-betrouwbaarheidsband geven voor de twee groepen. Elke betrouwbaarheidsinterval is zo opgesteld dat met een betrouwbaarheid van 95% gezegd kan worden dat elke waarde van een willekeurige stap uit die groep op dat punt in het interval zal liggen.

De tweede methode die geanalyseerd is is de Student's t-test, dit is een veel voorkomende test in de statistiek en is daarom ook meegenomen in dit artikel. Voor de tweezijdige t-test is op diezelfde 100 punten de student's t-test uitgevoerd en na een multipliciteit correctie kan de conclusie getrokken worden.

Beide testen zullen de nulhypothese – de trajecten zijn identiek – verwerpen als er op tenminste een van de 100 losse testen die naast elkaar uitgevoerd worden een verschil te detecteren is. Als de trajecten op een punt verschillen kan daaruit direct afgeleid worden dat de volledige trajecten niet identiek zijn.

Multipliciteit correctie

Bij elke statistische test die uitgevoerd wordt is er een kans dat de nulhypothese onterecht verworpen wordt: α. Deze kans wordt gecontroleerd door het betrouwbaarheids level, (1 – α)100%. In alle methodes is gerekend met een alpha van 0,05 en dus een betrouwbaarheids level van 95%. Dit betekent dat als een test 100 keer uitgevoerd wordt, in maximaal 5% van de gevallen de test ten onrechte de nulhypothese verwerpt.

Voor de testen waarbij gelijktijdig 100 testen worden uitgevoerd moeten we verwachten dat er in totaal 5 valse verwerpingen zijn. Hiervoor worden de individuele alpha's aangepast met behulp van een multipliciteit correctie (Dudoit, Shaffer & Boldrick, 2003), in dit geval Bonferroni correctie. De gecorrigeerde α voor de individuele punten is dan $\alpha_i = \alpha/s$, met s het aantal individuele punten, 100, waarop de test wordt uitgevoerd. Hierdoor blijft het betrouwbaarheids level van de gecombineerde test op 95%.

Bij de laatste test wordt er direct op het hele traject getest en is er dus ook geen multipliciteit correctie nodig. Met behulp van bootstrapping wordt de kritieke waarde, de drempelwaarde voor statistische significantie, bepaald door middel van pseudo observaties. Deze worden gecreëerd door een random residufunctie, op te tellen bij het traject. De residufunctie is het verschil tussen een van de 60 trajecten en het gemiddelde traject. Dit wordt gedaan voor alle 60 trajecten en zo wordt er een nieuwe pseudo observatie gecreëerd. Bij bootstrapping wordt dit proces heel vaak herhaald, in dit geval 1000 keer, en zo wordt er een pseudo dataset gecreëerd.

Voor elke pseudo observatie is de pseudo teststatistiek S_N^+ berekend volgens

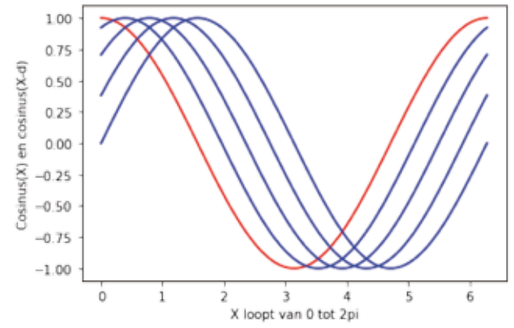
$$S_N^+ = \frac{n_1 n_2}{N} \left| \overline{X}_{1,n_1} - \overline{X}_{2,n_2} \right|^2$$

en de kritieke waarde, C_α van deze test is geschat door de 1 – α percentiel te nemen van de gesorteerde lijst met deze pseudo observaties (Paparoditis & Sapatinas, 2016).

Om uiteindelijk een conclusie te kunnen trekken wordt de teststatistiek van de originele dataset vergelijken met de kritieke waarde, en zal de nulhypothese verworpen worden als de teststatistiek groter is dan de kritieke waarde. In dit geval is het verschil tussen de groepen zo groot, dat dit niet door willekeurige ruis zal zijn.

Resultaten

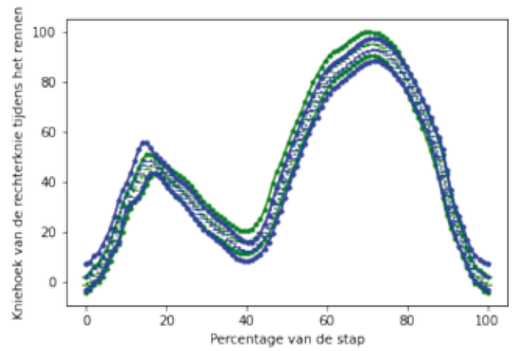
Elke methode is eerst getest op kunstmatige data zodat de validiteit van de testen kon worden geschat. Om zulke kunstmatige data te verkrijgen wordt een echt stap patroon en algemene ruis gekozen. Hiervoor is een cosinus en random normaal verdeelde ruis gebruikt. De methoden waren gevalideerd zodra de testen op kunstmatige data gecontroleerd waren door alpha en het betrouwbaarheidsniveau. Voor elke methode is de power geschat, dit is de kans dat de test correct de nulhypotheses verwerpt. De power is ingeschat door te testen op de rode groep met een blauwe groep uit figuur 5 waarbij de blauwe groepen lopen van cosinus tot sinus, waarbij het verschil tussen de groepen dus steeds duidelijker wordt. Zoals in tabel 1 gezien kan worden is de test met de bootstrap het best om kleine verschillen te detecteren



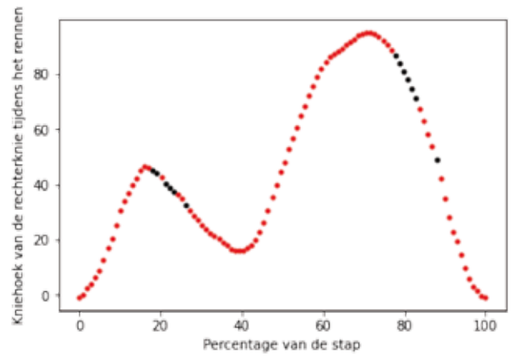
Figuur 5. Power

δ	% verwerpingen A betrouwbaarheidsintervallen	% verwerpingen B student's t-test	% verwerpingen C bootstrap test
0	0	4	4
$\frac{\pi}{200}$	0	7	8
$\frac{2\pi}{200}$	0	18	46
$\frac{3\pi}{200}$	0	34	96
$\frac{4\pi}{200}$	0	62	100
$\frac{5\pi}{200}$	0	86	100
$\frac{6\pi}{200}$	0	100	100
⋮	⋮	⋮	⋮
$\frac{3\pi}{10}$	0	100	100
$\frac{4\pi}{10}$	11	100	100
$\frac{\pi}{2}$	96	100	100

Tabel 1



Figuur 6. Convidence bands



Figuur 7. T-test

en heeft daardoor dus de hoogste power.

Tot slot zijn de methoden uitgevoerd met de twee groepen data van het vermoeidheidsprotocol. De resultaten hiervan zijn te zien in de figuren 6 en 7. In figuur 6 is methode A door middel van betrouwbaarheidsintervallen te zien. Volgens deze methode zullen twee gemiddelde trajecten verschillend zijn omdat de betrouwbaarheids banden overlap hebben op alle punten. Figuur 7 toont methode B. De tweezijdige t-test verwerpt de nulhypothese wel; er zijn in figuur 7 meerdere punten te vinden waarop de trajecten niet identiek zijn. Voor methode C. Bootstrap test S_N = 24843,21 is groter dan C_α = 598,87. Dus de nulhypothese wordt verworpen. Dit betekent dat, volgens deze testen, het erg onwaarschijnlijk is dat de verschillen komen door willekeurige ruis, en dus komen door het verschil, de vermoeidheid, in de twee groepen.

Conclusie

Over het algemeen blijkt een test door middel van betrouwbaarheidsintervallen niet geschikt om de bewegingen van de kniehoeken te analyseren, deze is namelijk niet in staat om zulke kleine verschillen te detecteren. Daarentegen zijn de tweezijdige t-test en de bootstrap test dit wel. De power van de bootstrap test bleek het grootst. Daarom is de bootstrap test het best om te gebruiken voor verder onderzoek richting blessurepreventie tijdens het hardlopen.

Literatuur

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science, 18*(1), 71–103. http://www.jstor.org/stable/3182872

Mechelen, W. van. (1992). Running injuries: A review of the epidemiological literature. *Sports Medicine, 14*(5), 320–335.

Paparoditis, E., & Sapatinas, T. (2016). Bootstrap-based K-sample testing for functional data. *arXiv* 1409.4317.

Bolten, M. M. (2021). *Detection of changes in movement patterns of runners*. http://essay.utwente.nl/86782/

Dankwoord van Marjolein Bolten

Dit artikel is een samenvatting van mijn bacheloropdracht voor de bachelor Applied Mathematics bij het project Sports, Data & Interaction, opgericht door de faculteit EEMCS van de Universiteit van Twente (een samenwerking van Applied Mathematics, Technical Computer Science and Electrical Engineering). Ik werkte samen met dr. Katharina Proksch en Rupsa Basu MSc, die ik ook wil bedanken voor hun hulp tijdens het schrijven van dit artikel. Ook wil ik prof. dr. Nico van Dijk bedanken voor het proeflezen en de suggesties bij het schrijven van dit artikel.

E-mail: m.m.bolten@student.utwente.nl