Universiteit
Leiden
The Netherlands

Permutation-based inference for high-dimensional data.

Hemerik, J.

Citation
Hemerik, J. (2018, May 1). *Permutation-based inference for high-dimensional data*. Retrieved from https://hdl.handle.net/1887/61825

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/61825

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/61825 holds various files of this Leiden University dissertation.

Author: Hemerik, J.
Title: Permutation-based inference for high-dimensional data
Issue Date: 2018-05-01

Samenvatting

Wetenschappelijk onderzoekers zijn vaak geïnteresseerd in een nulhypothese over een populatie. Een nulhypothese is een uitspraak zoals: "Rokers hebben (gemiddeld) dezelfde bloeddruk als niet-rokers."Een ander voorbeeld is de hypothese dat bij Europese vrouwen met een bepaalde ziekte, een nieuw medicijn A de ziekte (gemiddeld) niet beter verhelpt dan het gangbare medicijn B. In het eerste voorbeeld is de populatie die onderzocht wordt 'alle mensen', in het tweede geval 'alle Europese vrouwen met de ziekte.'

Om de eerste hypothese te toetsen, is het geen doen om van alle Nederlanders de bloedruk te meten. Dit zou bijvoorbeeld te veel tijd en geld kosten. In het algemeen meten onderzoekers dan ook niet de hele populatie waarin ze geïnteresseerd zijn, maar een representatieve steekproef. Bij het tweede voorbeeld kan bijvoorbeeld een groep van 20 vrouwen verzameld worden. Hiervan krijgen 10 willekeurig gekozen vrouwen medicijn A en de 10 andere vrouwen medicijn B. De vrouwen krijgen niet te horen welk medicijn ze hebben gekregen. Als bij deze vrouwen medicijn A steeds veel beter werkt dan medicijn B, is er bewijs dat A beter is dan B. We kunnen dan overwegen om de hypothese te verwerpen, dat A niet beter werkt dan B.

Wat het analyseren van zulke steekproeven zo interessant maakt, is dat we nooit met zekerheid kunnen zeggen dat A beter is dan B. Het kan namelijk zo zijn dat medicijn A door puur toeval aan mensen werd gegeven die een betere weerstand hadden. In het algemeen is er altijd een risico dat een hypothese onterecht verworpen wordt, doordat een steekproef door puur toeval een misleidend beeld geeft van de populatie. Er is dus altijd sprake van onzekerheid als conclusies over populaties worden gedaan op basis van steekproeven. De statistiek houdt zich bezig met het kwantificeren van die onzekerheid.

Het is belangrijk om hypotheses zodanig te toetsen, dat de kans op het verwerpen van een ware hypothese klein is. Dit zouden we kunnen we doen

door een hypothese alleen te verwerpen, als de kans dat deze waar is heel klein is, op basis van de steekproef. Het berekenen van die kans blijkt echter nogal problematisch te zijn, en in de (frequentische) statistiek probeert men in plaats daarvan te bepalen hoe onwaarschijnlijk de data zouden zijn, als de hypothese waar was.

De onwaarschijnlijkheid van de data onder de nulhypothese wordt gemeten met een getal, de *toetsstatistiek*. Hoe onwaarschijnlijker de data zijn onder de nulhypothese, hoe groter de toetsstatistiek is. Als de toetsstatistiek onwaarschijnlijk groot is onder de nulhypothese, wordt de nulhypothese verworpen. De onwaarschijnlijkheid van de toetsstatistiek wordt gemeten met de p-waarde, een getal tussen 0 en 1. Hoe groter de toetsstatistiek is, hoe kleiner de p-waarde is. Als de nulhypothese waar is, is de p-waarde (bij benadering) uniform verdeeld tussen 0 en 1. Vaak verwerpt men de nulhypothese als de p-waarde 0.05 of lager is. Is de p-waarde groter dan 0.05, dan zegt men dat er niet voldoende bewijs is om de hypothese te verwerpen.

Om te begrijpen waarom dit een goede manier van toetsen is, kunnen we ons realiseren dat er twee soorten fouten kunnen worden gemaakt. Ten eerste kunnen we een hypothese verwerpen, terwijl deze in werkelijkheid waar is. Dit heet een type-I fout. Ten tweede kunnen we een hypothese niet-verwerpen als deze in werkelijkheid onwaar is. Dit heet een type-II fout. Meestal worden type-I fouten erger gevonden dan type-II fouten.

Dat een type-I fout ernstige gevolgen kan hebben, is duidelijk. Stel bijvoorbeeld dat we concluderen dat medicijn A beter werkt dan medicijn B, terwijl medicijn A in werkelijkheid minder goed werkt. Dan zal, op basis van dit resultaat, in de toekomst misschien medicijn A – het verkeerde medicijn – aan patiënten gegeven worden.

Een type-II fout wordt vaak minder erg gevonden. Eén van de redenen is dat bij het niet-verwerpen van een hypothese eigenlijk geen uitspraak wordt gedaan, dus ook geen foutieve. Een hiermee samenhangende reden is dat de hypothese meestal zegt dat ‘er niks interessants aan de hand is’ (bijvoorbeeld: ‘A werkt niet beter dan het gebruikelijke medicijn B’), waardoor er geen nadruk wordt gelegd op niet-verworpen hypotheses in wetenschappelijke publicaties. Sterker nog, niet-verworpen hypotheses worden vaak überhaupt niet vermeld in publicaties (zogeheten ‘publication bias’). Type-II fouten zijn dus jammer, maar vervuilen de wetenschappelijke literatuur niet zo erg als type-I fouten.

In het voorbeeld van medicijnen bijvoorbeeld, betekent een type-II fout dat medicijn A beter werkt, maar dit niet geconcludeerd wordt. Dit is jammer, maar er wordt geen (foute) conclusie getrokken. Bovendien krijgt het niet-verwerpen van de hypothese weinig aandacht in de wetenschappelijke gemeenschap. Daardoor zal A in de toekomst waarschijnlijk nog een keer met B vergeleken worden, wat dan wel tot verwerping van de hypothese kan leiden.

Aangezien men de nulhypothese verwerpt als de data onwaarschijnlijk zijn onder de nulhypothese, is de kans op verwerpen klein als de hypothese waar is. Er wordt dus expliciet voor gezorgd dat de kans op type-I fouten klein blijft: precies wat men wil.

Dit proefschrift legt voor een belangrijk deel de nadruk op *multiple testing*, oftewel meervoudig toetsen. Dit is het toetsen van veel (strict gesproken: twee of meer) hypotheses tegelijk. We zouden bijvoorbeeld voor 20,000 genen de hypothese kunnen toetsen, dat de activiteit van het gen geen verband houdt met een bepaald fenotype (bijvoorbeeld de variant van kanker, die iemand heeft). Bij zoveel hypotheses moet de statisticus, op basis van de hem beschikbaar gestelde data, voor elke hypothese bepalen of deze verworpen wordt.

Als er veel hypotheses getoetst worden, kunnen er veel type-I fouten gemaakt worden. Als de onderzoeker er vrij zeker van wil zijn dat er geen enkele type-I fout gemaakt wordt, moet hij ervoor zorgen dat een hypothese alleen verworpen wordt, als er extreem veel bewijs tegen is. Het gevolg van zulke strenge eisen is vaak dat er maar weinig hypotheses verworpen worden, zelfs als in werkelijkheid veel van de hypotheses onwaar zijn.

Soms gaat het er niet om, ervoor te waken dat er helemaal geen type-I fouten zijn, maar wil de onderzoeker dat met grote kans bijvoorbeeld 95% van de verworpen hypotheses onwaar zijn. Hoofdstuk 3 en 4 van dit proefschrift bevatten nieuwe statistische methoden die dit kunnen garanderen.

In hoodstuk 3 wordt een bestaande *multiple testing* methode verbeterd, getiteld SAM ("Significance Analysis of Mircorarrays"). Deze methode werd in eerste instantie gebruikt voor genetische data, maar is breder toepasbaar. De oorspronkelijke SAM methodologie schat de fractie van type-I fouten onder alle verworpen hypothesen. Als de door SAM geschatte fractie type-I fouten laag is, suggereert dit dat veel van de verworpen hypothesen onwaar zijn.

De door SAM geschatte fractie van type-I fouten kan ver van de

werkelijke fractie af liggen. In hoofdstuk 3 wordt een methode ontwikkeld om een bovengrens voor de werkelijke fractie te geven. De gebruiker kan zelf kiezen hoe betrouwbaar deze bovengrens moet zijn. Op deze manier kan de gebruiker bijvoorbeeld een bovengrens berekenen die boven de ware fractie type-I fouten ligt met een kans van tenminste 95%. Als de gevonden bovengrens bijvoorbeeld 0.15 is, dan weet de gebruiker met 95% zekerheid dat de fractie type-I fouten niet groter is dan 0.15.

De methode van hoodstuk 3 is geïmplementeerd in een klein software pakket, dat online vrij beschikbaar is. Hierdoor kan de methode gemakkelijk door statistici wereldwijd gebruikt worden. De appendix van dit proefschrift bevat een gebruiksaanwijzing voor deze software.

Hoofdstuk 4 is gerelateerd aan hoofdstuk 3. Ook hier worden bovengrenzen gegeven voor de fractie van type-I fouten in een collectie verworpen hypothesen. In tegenstelling tot hoofdstuk 3 echter, zijn de bovengrenzen in hoofdstuk 4 nog steeds betrouwbaar als het verwerpingscriterium wordt bepaald op basis van de data. Hierdoor krijgt de gebruiker meer vrijheid in het aanpassen van het verwerpingscriterium na het kijken naar de data.

Statistiek gaat niet alleen over toetsen, maar ook over het maken van voorspellingsmodellen. (Hierbij staat het kwantificeren van onzekerheid nog steeds heel centraal.) Het kan bijvoorbeeld zo zijn dat het per persoon verschilt welk medicijn waarschijnlijk het beste zal werken. Dat kan afhangen van variabelen zoals geslacht en leeftijd. Het kan dan nuttig zijn om een statistisch model te maken, dat voorspelt welk medicijn het beste bij iemand werkt, gegeven eigenschappen als leeftijd, geslacht en genen.

Hoofdstuk 5 van dit proefschrift gaat over het toetsen binnen zulke modellen. Het gaat er hierbij om aan te tonen of bepaalde variabele, bijvoorbeeld leeftijd, een statistisch aantoonbare invloed op de uitkomstvariabele (bijvoorbeeld het optimale medicijn) heeft.

Zulke statistische toetsen worden vaak uitgevoerd onder bepaalde aannames, bijvoorbeeld dat de variantie van de uitkomstvariabele op een bepaalde manier afhangt van de verwachtingswaarde. In hoofdstuk 5 wordt een nieuwe toets geconstrueerd, die vaak nog goed werkt als niet aan de aannames voldaan is. Daardoor kan de gebruiker er meer op vertrouwen dat de kans op type-I fouten echt onder (zeg) 5% ligt, zelfs als bepaalde aannames niet helemaal kloppen. Net zoals de methode van hoofdstuk 3, is de methode van hoofdstuk 5 geïmplementeerd in software die online vrij beschikbaar is.

De toets in hoofdstuk 5 is gerelateerd aan de *multiple testing* methodes in de eerdere hoofstukken. Daardoor wordt het mogelijk om die methodes toe te passen in de modellen waar hoofdstuk 5 over gaat.