Samenvatting

Het opstellen en toetsen van statistische hypotheses neemt van oudsher een zeer belangrijke plaats in binnen bijna elke vorm van wetenschappelijk onderzoek. Het doel van statistisch hypothese toetsen is om op basis van steekproefgegevens een bepaald effect in *de populatie als geheel* aan te tonen. Doorgaans wordt ervoor gekozen om als uitgangspunt te nemen dat het effect afwezig is, waarna er wordt gekeken of dit uitgangspunt houdbaar blijft in het licht van de gevonden resultaten. De hypothese die getoetst wordt is dus de hypothese dat er geen effect is, de zogeheten nulhypothese, en de hoop is doorgaans dat deze hypothese verworpen kan worden.

Er zijn heel veel verschillende nulhypotheses te bedenken maar de nulhypotheses die binnen dit proefschrift centraal staan zijn gerelateerd aan het biologische fenomeen van genexpressie. Een gen is niets meer dan een afgebakend stukje van ons DNA, en wanneer een gen tot expressie komt betekent dit dat dit gen wordt gekopieerd naar messenger-RNA dat vervolgens weer kan worden vertaald naar een specifiek eiwit. Een gen dat "tot expressie komt" resulteert dus in de synthese van eiwitten. Deze eiwitten kunnen allerhande functies hebben binnen ons lichaam en de variatie in genexpressie is wat een cel zijn karakter of functie geeft. Tegenwoordig bestaan er technieken waarmee voor duizenden genen tegelijk de mate van expressie kan worden gemeten. Wanneer op deze manier de genexpressie voor zowel gezonde als zieke personen in kaart wordt gebracht, kan er vervolgens gekeken worden of er verschillen in de genexpressie aan te wijzen zijn tussen deze groepen.

Om voor één gen te kijken of het een rol speelt in het al dan niet hebben van een bepaalde ziekte, kan de nulhypothese worden getoetst dat de expressie van dit gen niet verschilt tussen gezonde mensen en mensen die lijden aan deze ziekte. Omdat we niet de genexpressie kunnen meten bij alle gezonde mensen en doorgaans ook niet bij alle mensen die aan een bepaalde ziekte lijden, zullen we onze uitspraak over deze groepen, de zogeheten populatie, moeten doen op basis van meetgegevens binnen een selecte groep, de zogeheten steekproef. Het verschil in genexpressie tussen de zieke en gezonde mensen in de steekproef levert een schatting op voor het echte verschil zoals aanwezig in de gehele populatie. Hoe groter de steekproef is en hoe minder variatie in genexpressie er is binnen de groep zieke respectievelijk gezonde mensen, hoe betrouwbaarder deze schatting is. Op basis van deze schatting en de betrouwbaarheid ervan wordt nu een uitspraak gedaan over de verwachte waarde van het echte verschil en op basis hiervan kan de nulhypothese al dan niet verworpen worden.

Nu kan het *bij toeval* gebeuren dat we binnen de steekproef een duidelijk verschil in genexpressie vinden tussen de gezonde en zieke mensen terwijl dat verschil er in de populatie als geheel niet is. In dat geval zouden we ten onrechte kunnen concluderen dat de nulhypothese verworpen mag worden, wat zou leiden tot het maken van een *type I fout*. Ook het omgekeerde kan gebeuren, namelijk het ten onrechte niet verwerpen van de nulhypothese, omdat we in de steekproef geen verschil zien, terwijl dit verschil er in werkelijkheid wel is. In dat geval zou een *type II fout* worden gemaakt. Beide fouten zijn niet geheel te voorkomen, maar het risico om een verkeerde uitspraak te doen kan wel onder controle gehouden worden. In dit proefschrift ligt de nadruk op het voorkomen van type I fouten, dus het voorkomen van onterechte verwerpingen van de nulhypothese.

Bij het toetsen van één nulhypothese wordt voor het maken van een type I fout doorgaans een foutmarge van 5% gehanteerd. De kans dat de nulhypothese onterecht verworpen wordt is dan dus maximaal 5%. Zolang het bij het toetsen van één hypothese blijft, is deze kans goed te verantwoorden. In dit proefschrift ligt de nadruk echter niet op het toetsen van één hypothese, maar op het toetsen van verschillende hypotheses tegelijkertijd. Stel dat we geïnteresseerd zijn in het gedrag van 1000 genen. Dit geeft 1000 nulhypotheses. Stel dat alle nulhypotheses waar zijn en dat deze 1000 genen in werkelijkheid niets te maken hebben met het al dan niet ziek zijn. Als we nu per nulhypothese een kans van 5% hebben om de hypothese onterecht te verwerpen, verwachten we 50 nulhypotheses te verwerpen en daarmee dus 50 "belangrijke genen" aan te wijzen terwijl deze in werkelijkheid niets met de ziekte van doen hebben. Om in het geval van meerdere nulhypotheses het aantal onterechte verwerpingen laag te houden, kunnen we dus niet langer elke hypothese toetsen alsof dit de enige hypothese is.

Om deze reden zijn er methodes ontwikkeld die de "familywise error rate" (FWER) controleren. De FWER is gedefinieerd als "de kans op minstens één foutieve verwerping" en als de FWER onder bijvoorbeeld 0.05 gehouden kan worden, betekent dit dat de kans op minstens één type I fout onder de 5% blijft. Een simpele manier om dit te bewerkstelligen is bijvoorbeeld door voor elke afzonderlijke toets nog maar een foutmarge van 5% gedeeld door het totale aantal toetsen te hanteren. In het voorbeeld met 1000 hypotheses zou dit dus neerkomen op een foutmarge van 0.005% per toets. Het nadeel van deze manier is echter dat een kleinere foutmarge er ook toe leidt dat het moeilijker wordt om een *onware* nulhypothese te verwerpen, terwijl dat juist is wat we zouden willen doen. Er zijn echter ook andere manieren om de kans op één of meer fouten te controleren zonder de foutmarge per afzonderlijke toets zo drastisch te verlagen en in dit proefschrift worden een aantal nieuwe methodes geïntroduceerd die precies dit doen.

Wat alle nieuwe methodes met elkaar gemeen hebben is dat er niet alleen wordt gekeken naar nulhypotheses die betrekking hebben op één specifiek gen, maar ook naar nulhypotheses die betrekking hebben op groepjes genen. Deze nulhypotheses stellen dat de genexpressie van *alle genen binnen een groep* gelijk zijn tussen de groep zieke en gezonde mensen. Het interessante aan dergelijke "geagreggeerde" nulhypotheses is dat er soms effecten zichtbaar gemaakt kunnen worden die op het niveau van de losse genen niet zichtbaar zijn. Als verschillende genen samenwerken en zich allemaal net wat anders ge-

dragen tussen zieke en gezonde mensen is dat op het individuele genniveau vaak moeilijk te zien terwijl het op een geaggregeerd niveau wel aantoonbaar is.

Een ander voordeel aan het toetsen van geaggregeerde hypotheses is dat er logische verbanden kunnen ontstaan tussen de hypotheses die vervolgens gebruikt kunnen worden in de toetsprocedure. Als we bijvoorbeeld een nulhypothese hebben verworpen voor een groepje van 2 genen, zeg gen A en gen B, en we ervan uitgaan dat dit een terechte verwerping was, weten we dat minstens één van deze twee genen verband houdt met het al dan niet ziek zijn. Als we nu de losse nulhypotheses voor alleen gen A en alleen gen B toetsen, weten we dat maar maximaal één van deze twee hypotheses waar kan zijn en deze informatie kunnen we gebruiken bij het vaststellen van de individueel toegestane foutmarges. Normaliter zou je de individuele foutmarges moeten corrigeren voor het feit dat we twee toetsen uitvoeren, maar omdat er maar maximaal één ware nulhypothese tussen zit en we dus maar maximaal één type I fout kunnen maken, hoeft de foutmarge in dit geval niet gecorrigeerd te worden. Dit type redenering speelt een belangrijke rol in dit proefschrift.

Een tweede manier waarop de logische verbanden gebruikt kunnen worden is voor het doen van uitspraken als: "binnen deze groep genen zijn minimaal 10 genen geassocieerd met de ziekte". Zoals al eerder opgemerkt kan het voorkomen dat een groep genen wel met een ziekte in verband kan worden gebracht, terwijl dat niet hoeft te gelden voor individuele genen binnen deze groep. Toch kan men nu zeggen dat er minstens één individueel gen binnen die groep moet zijn dat met de ziekte te maken heeft. Als er verschillende genen uit verschillende relevante groepen worden samengenomen, kunnen er vervolgens uitspraken gedaan worden over het minimale aantal aan de ziekte gerelateerde genen binnen deze nieuwe groep. De mogelijkheid tot het doen van dergelijke uitspraken geeft onderzoekers de ruimte om post-hoc (dus als alle hypotheses getoetst zijn) groepen genen te selecteren waarvan vervolgens vastgesteld kan worden hoeveel genen binnen deze groep minimaal geassocieerd zijn met de onderzochte ziekte. Door groepen zo te kiezen dat er zoveel mogelijk relevante genen in vallen, kan zo bijvoorbeeld worden bepaald naar welke groepjes genen men specifieker wil kijken in een vervolgexperiment. Bij alle methodes die worden geïntroduceerd in dit proefschrift is er de mogelijkheid tot het doen van dit type post-hoc analyse.

In hoofdstuk 2 tot en met 5 van dit proefschrift worden vier verschillende methodes beschreven om nulhypotheses gerelateerd aan specifieke genen en aan groepen genen te toetsen. Hoewel de methodes ook in andere contexten gebruikt kunnen worden, zullen we daar in deze samenvatting niet verder op ingaan. Wat alle methodes gemeen hebben is dat ze gebruik maken van de aanwezige logische verbanden tussen de nulhypotheses om zoveel mogelijk verwerpingen te kunnen realiseren, terwijl er tegelijkertijd voor wordt gezorgd dat de kans op één of meer onterechte verwerpingen begrensd blijft. Een andere overeenkomst is de mogelijkheid tot het doen van de zojuist beschreven post-hoc analyses. Een laatste overkoepelende factor is dat bij alle methodes de nadruk is gelegd op een efficiënte implementatie om de methodes te kunnen gebruiken voor een zo groot mogelijk aantal hypotheses. Een belangrijk verschil tussen de vier methodes is de manier waarop

de groepen genen die getoetst worden samengesteld mogen worden en de volgorde waarin deze groepen worden getoetst.

In hoofdstuk 2 wordt een methode beschreven waarin voor een verzameling van n genen, met n een willekeurig getal, zowel alle losse genen als alle denkbare groepen genen getoetst worden op hun relevantie. Omdat er geen restrictie is op de samenstelling van de groepen, zijn er heel veel (namelijk $2^n - 1$) nulhypotheses die getoetst moeten worden. Dit kan computationeel gezien alleen maar als er wel een restrictie wordt gezet op het type toets dat wordt gebruikt. In hoofdstuk 2 gebruiken we als specifieke toets een "Simes toets" en laten we zien hoe we toetsuitkomsten (wel/niet verwerpen) van bepaalde slim gekozen nulhypotheses kunnen gebruiken om ervoor te zorgen dat we niet alle toetsen daadwerkelijk uit hoeven te voeren, terwijl we toch voor elke hypothese weten of deze wel of niet verworpen kan worden. Ook laten we zien dat de resultaten van onze methode, toegespitst op de nulhypotheses die horen bij specifieke genen (en dus niet bij groepjes genen), precies overeenkomen met de resultaten van een veelgebruikte, reeds bestaande methode. Onze manier van uitrekenen zorgt er echter voor dat de resultaten vele malen sneller gevonden worden, waardoor het mogelijk wordt een veel groter aantal nulhypotheses te toetsen dan voorheen.

In hoofdstuk 3, 4 en 5 kiezen we voor een andere aanpak. In deze hoofdstukken wordt geen restrictie op het soort toets gelegd, maar wel op het type nulhypothese, dus op de specifieke samenstelling van de te toetsen groepen genen. Omdat we, wanneer de toets vrij te kiezen is, niet langer alle groepen kunnen bekijken, hebben we methodes bedacht die biologisch interessante groepen genen toetsen. In hoofdstuk 3 kijken we naar groepen "aangrenzende" genen, dus naar genen die fysiek gezien bij elkaar in de buurt liggen. In eerste instantie wordt het grootste aangrenzende gebied getoetst, en als er aanwijzingen zijn dat er binnen dit gebied genen zijn die geassocieerd zijn met de ziekte, toetsen we steeds kleinere gebiedjes om zo nauwkeurig mogelijk te kunnen bepalen welke specifieke genen een rol spelen bij dit ziektebeeld. In hoofdstuk 4 en 5 kijken we niet specifiek naar aangrenzende genen, maar worden de groepen genen van tevoren samengesteld op basis van biologische kennis. Voor heel veel processen binnen ons lichaam is al goed bekend welke groepen genen daarin een rol spelen, en deze groepen zijn dus interessant om te bekijken. Het verschil tussen de in hoofdstuk 4 en 5 gepresenteerde methodes is de volgorde waarin de groepen genen getoetst worden. De methode uit hoofdstuk 4 toetst een groep genen pas op het moment dat alle grotere groepen al getoetst en verworpen zijn en er dus aanleiding is om binnen deze groep genen verder te zoeken naar de specifieke genen die dit effect veroorzaken. De methode uit hoofdstuk 5 hanteert daarentegen geen specifieke volgorde tijdens het toetsen en toetst alle vooraf gespecificeerde groepen genen tegelijkertijd.

Waar de focus in het eerste deel van dit proefschrift ligt op het vinden van genen of groepen genen die verband houden met het al dan niet ziek zijn, wordt er nog geen aandacht besteed aan de vraag welk effect een bepaald gen heeft. Zodra men weet *dat* een bepaald gen geassocieerd is met ziek zijn, zal de volgende vraag doorgaans zijn *hoe* deze connectie werkt en of verhoogde expressie van het gen een lager of juist een hoger risico

op ziekte met zich meebrengt. Het mooist zou zijn om op basis van iemands genexpressie vast te kunnen stellen of deze persoon al dan niet ziek is, of om op basis van deze informatie bijvoorbeeld vast te kunnen stellen hoe iemand op een behandeling zal reageren. Om dergelijke vragen te kunnen beantwoorden kunnen voorspelmodellen gebruikt worden en het verband tussen hypothese toetsen en voorspelmodellen wordt verder uitgelegd in hoofdstuk 6.

In hoofdstuk 7 gaan we vervolgens dieper in op het opstellen en valideren van voorspelmodellen. Zoals er bij het toetsen van hypotheses een kans is dat er wel een effect is binnen de steekproef maar niet in de populatie als geheel, kan het bij het opstellen van een voorspelmodel gebeuren dat het model fantastisch voorspelt op de huidige dataset, terwijl het voor vergelijkbare datasets weinig tot geen voorspellende waarde heeft. Om te voorkomen dat een voorspelmodel te specifiek op één bepaalde steekproef is aangepast, moet er bij het opstellen van het model al rekening gehouden worden met de voorspelprestaties op vergelijkbare data (dus op de populatie als geheel). Een veelgebruikte, maar vaak langzame, methode om dit te bewerkstellingen is om bij het opstellen van het model al een deel van de data apart te houden, die vervolgens gebruikt kan worden om het model te valideren op een onafhankelijke dataset. Op deze methode wordt dieper ingegaan in hoofdstuk 7. In ditzelfde hoofdstuk presenteren we ook een manier om vergelijkbare resultaten te kunnen behalen in een kortere tijd.

Het overkoepelende thema van dit proefschrift is dus het voorkomen van toevalstreffers. Wanneer er heel veel genetische informatie beschikbaar is en wetenschappers alle mogelijke verbanden willen onderzoeken, is er altijd een kans dat er iets gevonden wordt dat in werkelijkheid niets blijkt te zijn. Om te voorkomen dat zulke toevalstreffers in de statistische en medische literatuur als waarheden worden gepresenteerd, moet er al tijdens het proces kritisch gekeken worden hoe de kans op zulke toevalstreffers onder een bepaalde acceptabele grens gehouden kan worden. Verschillende manieren om op een verantwoorde wijze met grote aantallen interessante hypotheses en modellen om te gaan, zijn in dit proefschrift beschreven.