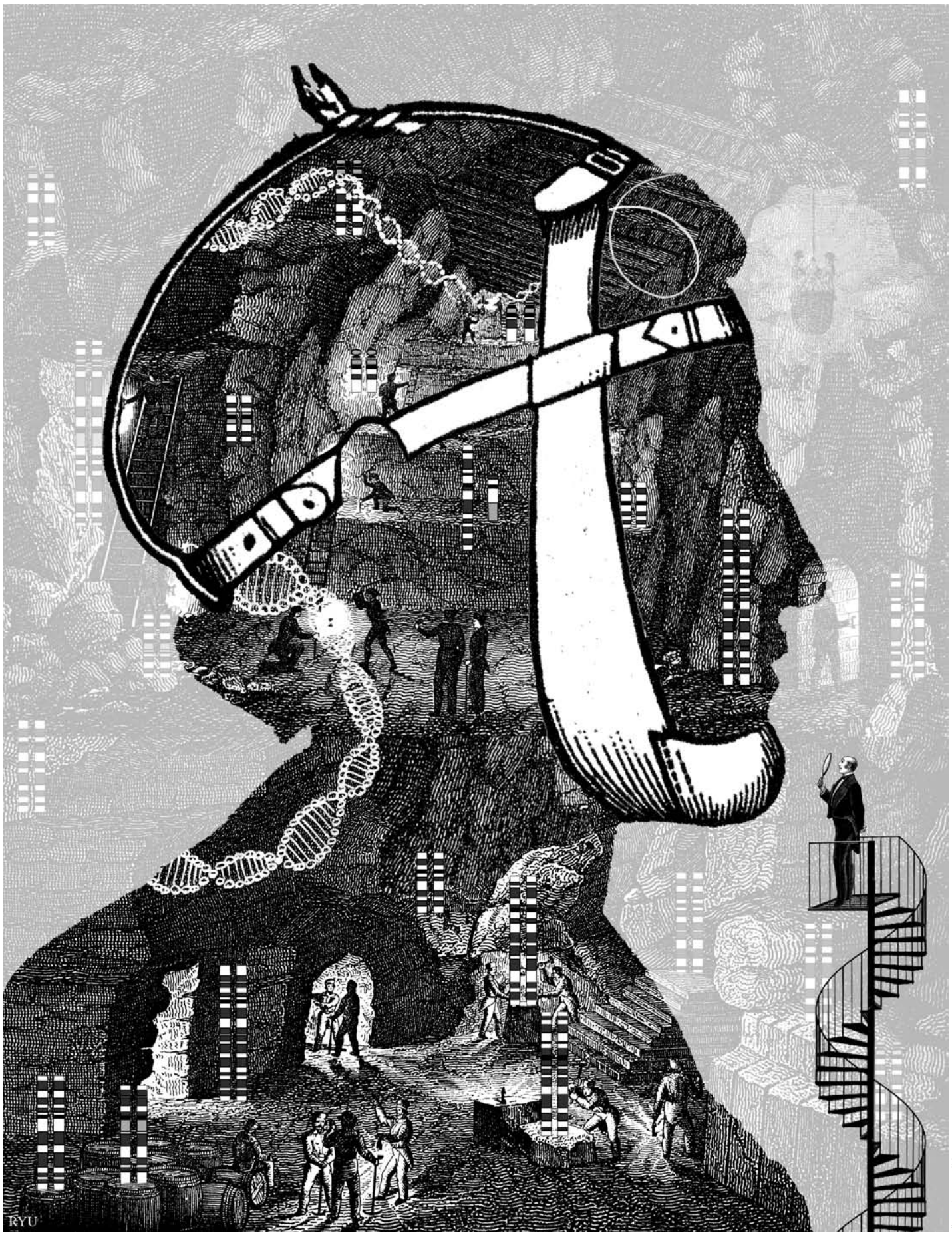


Illustratie: Ryu Tajiri

Mark van de Wiel
Afdeling Wiskunde, Vrije Universiteit Amsterdam, en
Afdeling Epidemiologie en Biostatistiek, VUmc, Amsterdam
mark.vdwiel@vumc.nl

Onderzoek

Statistiek op het genoom: 'Big Data', maar dan anders

In het voorjaar van 2014 werd Mark van de Wiel benoemd tot hoogleraar statistiek voor genoomanalyse aan de Faculteit Exacte Wetenschappen van de Vrije Universiteit en aan het VU Medisch Centrum. Dit artikel is gedeeltelijk gebaseerd op zijn oratie gehouden op 23 mei 2014. Hij beoogt duidelijk te maken waar de statistiek bijdraagt aan de analyse van Big Data, en dan in het bijzonder genomische data.

Big Data: Big p or Big n?

Big Data is overal tegenwoordig. De Googles en Facebooks van onze maatschappij verzamelen enorme hoeveelheden data op een automatische manier, onze vele mobiele telefoons leveren locatie-informatie voor bijvoorbeeld file-voorspellingen, en in een medische setting geven MRI-beelden en metingen aan het genoom gigantische aantallen datapunten per individu. Maar wat is 'Big'? Hiertoe moeten we eerst onderscheid maken tussen n: het aantal individuen en p: het aantal variabelen. En het is juist dat onderscheid dat zo essentieel is voor de statistische analyse van Big Data: als n groot is en veel groter dan p ($n >> p$) dan werkt traditionele statistiek goed, zelfs heel goed, vooral voor de vele asymptotische benaderingen die de statistiek rijk is. Wellicht de grootste uitdaging voor dit geval is het bestuderen van complexe, (niet-lineaire) modellen, omdat de grote n daartoe de ruimte geeft. Als echter $p >> n$ spreken we van hoog-dimensionale data, en stapelen de problemen zich op, omdat de traditionele statistiek niet meer werkt. Ik geef hier verderop enkele voorbeelden van. Genomische data zijn vrijwel altijd hoog-dimensionaal: niet zelden is p van de orde 10^5 terwijl $n \approx 100$. De oorzaak is simpel: nieuwe technologieën stellen moleculair biologen in staat met het grootste gemak veel stukjes van het genoom parallel te meten. Echter, zeker in klinische omgevingen, is het lastig en duur om weefsel te verkrijgen van grote aantallen (zieke) individuen. Ik vervolg met een voorbeeld van (al dan niet opzettelijke) verwarring tussen p en n in genomisch onderzoek.

Big p, Big Errors (p is geen n)

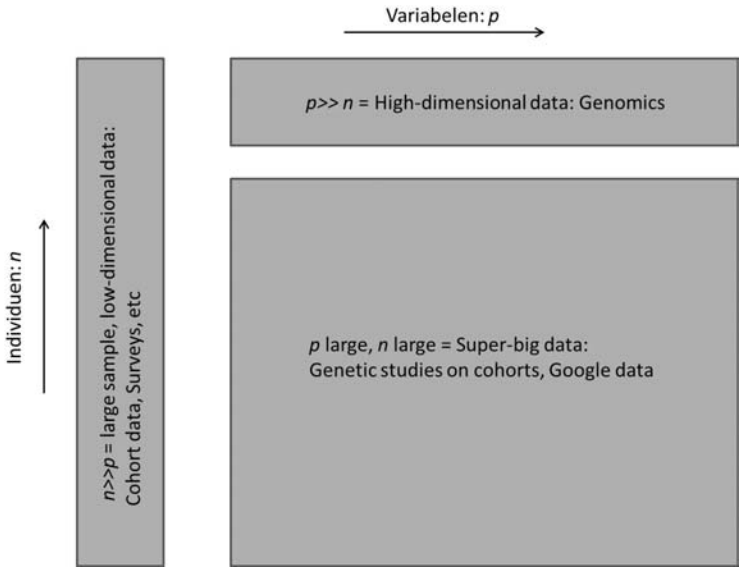
Er wordt wel gezegd: "There are lies, damned lies and statistics" [8]. Ik zou willen zeggen "wrong statistics". Met goede statistiek lieg je niet, sterker nog: met goede statistiek achterhaal je leugens, of op zijn minst foute conclusies. Een beroemd historisch voorbeeld in de genetica is het experiment van Mendel met de erwtenplanten [4]. Op dit experiment heeft hij zijn erfelijkheidswetten gebaseerd. Statisticus én geneticus Fisher gebruikte Pearsons chi-kwadraattoets om aan te tonen dat Mendels resultaten te goed om waar te zijn waren [2]. Wetenschappers zijn dus net mensen, en zelfs briljante wetenschappers als Mendel gaan weleens de mist in. Wanneer het genomisch onderzoek betreft, denk ik dat wetenschappelijke nalaatbaarheid een veel negatievere invloed heeft op de kwaliteit dan moedwillige fraude. Te vaak publiceren absolute toptijdschriften genomisch onderzoek waarvan de statistiek gewoon slecht is.

Een voorbeeld. Methylatie is een belangrijk proces op het DNA waarmee genen kunnen worden uitgeschakeld. Een artikel in *Nature* uit 2011 [3] beschrijft methylatie van gebieden op het genoom in stamcellen en normale cellen. De auteurs rapporteren vele gebieden die verschillend gemethyleerd zijn tussen deze twee types cellen. Prachtige plaatjes, krachtige conclusies. Het artikel is dan ook al meer dan 600 keer geciteerd. Ik was geïnteresseerd, want ik dacht dat men maar liefst tien biologische herhalingen had per conditie. Statistici zijn dol op herhalingen en tien is veel voor de dure techniek die werd gebruikt, 'bisulphite sequencing'.

Echter wat bleek: de herhalingen waren fictief. Men had het genoom in stukjes opgehakt en tien achtereenvolgende stukjes als onafhankelijke herhalingen beschouwd. Dus gedaan alsof dát de echte biologische variatie representeert. Dat is zoiets als zeggen dat de variatie tussen zoogdieren bestudeerd kan worden door de variatie tussen mensen te beschouwen. Voor het artikel leidde de sterk onderschatte variatie natuurlijk tot prachtige p-waardes. Deze waren blijkbaar klein genoeg om de referenten te overtuigen. Ik heb het aangekaart bij *Nature*; men vond het niet urgent genoeg voor publicatie van het commentaar, maar ik mocht het commentaar wel online plaatsen [6].

Big p, Big Problems

Waarin wijkt de statistiek voor $p >> n$ nu af van de traditionele statistiek? Simpel gezegd: in alle gevallen moet men een oplossing vinden voor een slecht-geconditioneerd probleem vanwege de 'curse of dimensionality'. Vergelijk het met het fitten van een



Figuur 1 Verschillende soorten 'Big Data'.

100ste-graads polynoom aan tien datapunten: daar zijn oneindig veel oplossingen voor. Ik noem een aantal gebieden waarop de laatste tien jaar veel onderzoek is gedaan.

Multiple testing

Hoe beperk je het aantal fout-positieven wanneer je meerdere hypotheses toetst? Bijvoorbeeld wanneer we willen weten welk gen zich anders gedraagt in een zieke groep dan in een gezonde groep individuen. Feitelijk een 'oud probleem', omdat ook al voor het $p >> n$-tijdperk men vaak meerdere statistische hypotheses toetste. De Bonferroni-regel [1] schrijft dan voor dat wanneer je de kans op minimaal één fout-positieve bevinding lager wilt hebben dan α, je simpelweg alleen variabelen (bijvoorbeeld genen) met een p-waarde kleiner dan α/p positief verklaart. Zie hier het dimensie-probleem: hoe groter p, des te kleiner de drempelwaarde. Veel onderzoek op het gebied van multiple testing richt zich óf op (a) andere, liberalere fout-positief criteria die meer meeschalen met p, zoals False Discovery Rate; óf op (b) het formuleren van geneste hypotheses en het dan gebruikmaken van de ontstane verbanden tussen de hypotheses om krachtigere procedures te ontwikkelen.

Geregulariseerde multiple regressie

Een klassiek probleem in de statistiek is het fitten van een lineaire multiple regressie met behulp van de kleinste-kwadratenmethode:

$$b = \text{argmin}_\beta \|Y - \mathbf{X}\beta\|_2,$$

waarbij Y de $n \times 1$-responsvector is (voor elk individu één respons), X de $n \times p$-matrix met verklarende variabelen (p voor elk individu) en β de $p \times 1$-vector met coëfficiënten. In een medische setting kan men bijvoorbeeld hiermee proberen de 'body mass index' (Y) te verklaren vanuit genetische factoren ($\mathbf{X}$). De oplossing is eenvoudig:

$$b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

Echter, als $p > n$ dan is $\mathbf{X}^T\mathbf{X}$ rang-deficiënt en dus niet inverteerbaar. Dit wordt opgelost door het minimalisatie probleem te regulariseren:

$$b = \text{argmin}_\beta \{\|Y - \mathbf{X}\beta\|_2 + \lambda\|\beta\|_q\},$$

met tuning parameter $\lambda > 0$ en veelgebruikte normen $q = 1$ (lasso-regressie) en $q = 2$ (ridge-regressie). Voor $q = 2$ kunnen we de oplossing direct opschrijven:

$$b = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T Y,$$

waarbij I de $p \times p$-identiteitsmatrix is. Het inverteren van de grote $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ matrix kunnen we efficiënt doen door singuliere-waardenontbinding van $\mathbf{X}$. Hoewel ridge-regressie vaak goede predictie-eigenschappen heeft (voorspel Y_{new} met $\mathbf{X}_{\text{new}}b$), selecteert het geen variabelen. Lasso-regressie doet dat wel: wanneer $q = 1$ zijn maximaal $k \leq n$ componenten van de p-dimensionale vector b ongelijk aan 0. Dit is prettig in de praktijk: een volgende keer hoeft de bioloog slechts k genen te meten. Onder een zogenaamde spaarzaamheids-aanname (het aantal echte niet-nulcoëfficiënten is klein), heeft de lasso goede asymptotische eigenschappen. Veel mathematisch statistisch werk richt zich op het verfijnen van deze resultaten voor (variaties op de) lasso. De lasso voorspelt echter niet zo goed als de ridge. Recentelijk is er veel aandacht voor een tweetal oplossingen: (i) hybriden tussen lasso en ridge, zoals de combinatie van de twee: het elastische net; of (ii) vanuit een ander perspectief: na ridge-regressie 'thresholding'-methodes toepassen om kleine coëfficiënten op nul te zetten zonder veel verlies van voorspellende waarde.

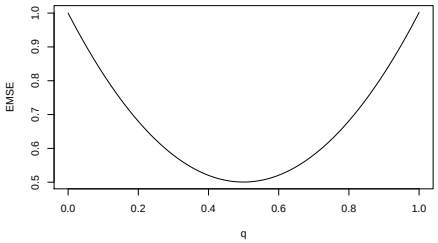
Moleculaire netwerken

Genen werken niet alleen, maar interacteren. Met elkaar, maar ook met andere moleculen zoals eiwitten. Wanneer we data hebben, kunnen we correlaties tussen genen berekenen om een idee te krijgen over welke interacties er mogelijk zijn. Dit is echter te naïef, omdat veel correlaties tussen twee genen veroorzaakt worden door interacties van beide genen met een of meerdere andere genen. Daarom is het beter om partiële correlaties te bestuderen, welke in het populaire Gaussische grafische model direct relateren aan de elementen van Σ^{-1} : de inverse correlatiematrix. Veel statistisch onderzoek op dit vlak richt zich dan ook op het schatten van Σ , een gruwelijk probleem, omdat het aantal onbekenden hierbij van de orde p^2 is. Weer spelen regularisatie-technieken zoals hierboven besproken een belangrijke rol. Het schatten van het netwerk is nu echter niet het eindpunt. Het geschatte netwerk bevat niet zelden honderden of duizenden verbindingen, wat tot een typisch 'haarbal'-plaatje leidt dat een bioloog natuurlijk onmogelijk kan interpreteren. Daarom richt men zich nu vaak op technieken om zulke netwerken samen te vatten, interessante kleine sub-modules te onttrekken, en ook om verschillen tussen netwerken (bijvoorbeeld gezond versus ziekte) te vinden.

Big p, Big Blessing

De grote p is niet altijd een vloek. Vaak is er veel structuur bekend over de p variabelen, in het bijzonder in genomische studies. Genen organiseren zich in padwegen, hetgeen modules zijn met één bepaalde biologische functie (bijvoorbeeld celdeling). Vaak zijn biologen geïnteresseerd of zo'n verzameling genen zich verschillend gedraagt onder twee of meer condities. Hiervoor zijn genset-toetsen ontwikkeld die in de regel meer onderscheidingsvermogen hebben dan toetsen op afzonderlijke genen. Zo kunnen we dus profiteren van de grote p. Een ander principe dat gebruik maakt van de grote p is 'Empirical Bayes', dat zich tevens goed leent voor het systematisch gebruikmaken van voorkennis.Het mooie van genomisch onderzoek is dat veel voorkennis opgedaan kan worden uit publiek beschikbare databases.

In het echte leven maken we continu gebruik van voorkennis om inschattingen te



Figuur 2 Expected mean square error (EMSE) als functie van q voor de schatter $m_i(q) = (1-q)X_i + q \cdot m$, wanneer $p = 1{,}000$.

doen. Een voorbeeld uit het casino, een plek waar een statisticus zich natuurlijk thuis voelt. Als iemand veel geld heeft verloren met een nieuw spel in het casino, en ik schat in dat die persoon redelijk intelligent is, dan speel ik liever het spel helemaal niet, dan dat ik eerst duizenden euro's uitgeef aan het spel om vervolgens waarschijnlijk te concluderen dat het voor mij ook niet werkt. Bayesiaanse methoden zijn bij uitstek geschikt hiervoor: de externe informatie kan verpakt worden in de zogenaamde prior, of à priori verdeling.

Kritiek uit de niet-Bayesiaanse hoek is dat in kleine, laag-dimensionale studies de keuze van de à priori verdeling subjectief is. Welnu, dat is het mooie aan hoog-dimensionale data: we hebben zoveel data van soortgelijke entiteiten, bijvoorbeeld genen, dat het mogelijk is om de à priori verdeling te schatten. Dat noemen we 'Empirical Bayes'. De te gebruiken bronnen van externe informatie zijn goeddeels keuzes, maar de relevantie van die bronnen wordt bepaald door de data zelf.

Laat ik even teruggaan naar het casinovoorbeeld, en ik neem aan dat het nieuwe spel een behendigheidsspel is, zoals bijvoorbeeld Black Jack. Stel nu dat niet alleen ik voor de keuze sta of en hoe vaak ik het spel ga spelen, maar u, de lezers, ook allemaal. Bovendien hebben we allemaal een ander, soortgelijk spel al heel vaak gespeeld. En helaas: we verloren daar gemiddeld gezien wat.

Nu zouden wij gelijk kunnen besluiten dit spel ook maar niet te spelen. Maar een slimmere strategie is de volgende: wij spelen allemaal het spel een paar keer met geringe inzet, waarna ieder voor zich gaat beslissen om door te gaan of niet. Waren we toen we dat andere spel speelden allemaal dronken, dan zullen we het nu met dit spel een stuk beter doen, hoop ik. En als 'dronken zijn' redelijkerwijs impliceert dat wij met dat vorige spel zomaar wat deden, dan zullen de data uitwijzen dat de correlatie tussen onze prestaties bij het nieuwe en oude spel zo goed als afwezig is. In dit geval vertellen de data ons dat we weinig gewicht moeten geven aan de externe informatie, hier onze prestaties bij het oude spel. Dan moet ieder voor zich een beslissing nemen op basis van de kleine hoeveelheid gegevens voor het nieuwe spel.

Het wordt echter interessanter als wij bij het vorige spel nuchter waren. En helemaal als de data van beide spellen ons dan vertellen dat er wél een behoorlijke correlatie is tussen de prestaties. Blijkbaar betreft het dan soortgelijke vaardigheden die nodig zijn om beide spellen relatief goed te kunnen spelen. Let wel: die correlatie, of algemener de à priori verdeling van een parameter die beschrijft in hoeverre de prestaties op elkaar lijken, kunnen we behoorlijk accuraat schatten (met Empirical Bayes technieken), want u bent (hopelijk) met velen. We kunnen dan veel meer gewicht toekennen aan de externe informatie. Juist omdat voor ieder van ons de data van het nieuwe spel van zeer beperkte omvang zijn, kan dit enorm helpen om betere beslissingen te nemen dan wanneer we deze informatie negeren. Niet voor iedereen zal de beslissing beter zijn, maar gemiddeld gezien vaak wel. En wellicht maken we als groep zelfs winst. Als we nu vooraf afspreken dat we die winst delen, dan is iedereen blij.

Een klein, iets meer wiskundig voorbeeld uit de genomics is het volgende. We veronderstellen een simpel model waarbij de (log)-genexpressie (= maat voor aantal kopieën van dat gen) voor gen i, X_i , een normale verdeling volgt met verwachtingswaarde μ_i en standaarddeviatie (sd) = 1. We meten echter vele genen en we veronderstellen dat μ_j een à priori normale verdeling volgt met verwachtingswaarde 0 en sd = 1 voor alle genen $j = 1, \ldots, p$. We willen μ_i schatten. De kwaliteit van een schatter m_i kwantificeren we met de zogenaamde Expected Mean Square Error (EMSE), de verwachtingswaarde onder de à priori verdeling van de MSE gedefinieerd als:

$$\text{MSE}(m_i) = [\text{Bias}(m_i)]^2 + \text{Variantie}(m_i).$$

Hier is Bias de verwachte afwijking van de schatter, m_i, van de waarheid, μ_i. Een simpele, zuivere (unbiased) schatter is $m_{i,1} = X_i$, met $\text{EMSE}(m_{i,1}) = 1$. Een mogelijke Empirical Bayes schatter (of krimp-schatter) is $m_{i,2} = X_i/2 + m/2$, waarbij m het gemiddelde van alle andere $p-1$ genexpressiemetingen is (dus we proberen te 'leren' van de andere genen). De schatter $m_{i,2}$ is niet zuiver, maar heeft wel een veel kleinere variantie. We kunnen de EMSE berekenen: $\text{EMSE}(m_{i,2}) = [1 + 1/(p-1)]/2$, welke dus bijna een factor 2 kleiner is dan die van $m_{i,1}$. Figuur 2 laat zien dat dit ook (nagenoeg) het minimum is. Dus ja, we kunnen echt leren van de andere genen. Let wel: als μ_i sterk verschilt van de à priori verwachtingswaarde van μ_j, $j \neq i$, dan kan de krimpschatter ook slechter zijn dan de simpele schatter. Voorzichtigheid is dus geboden als je vermoedt dat gen i een heel afwijkend gen is.

Big p, Big Business

Tot slot: Big Data, en dus ook hoog-dimensionale data, is Big Business. Dat is op zich goed nieuws voor de statistiek, welke niet voor niets door een hooggeplaatste Google-econoom de "sexiest job of the 21st century" [7] werd genoemd. Geen ander 'data-vakgebied' (bioinformatica, data science, et cetera) heeft zijn wortels zo goed verankerd in de wiskunde, waarmee we de mogelijkheden, maar juist ook de onmogelijkheden van veel hoog-dimensionale problemen kunnen aantonen. Dat laatste maakt ons gemiddeld pessimistischer dan andere data-broeders, en daarom ook minder populair bij (veel) biologen. Gelukkig is er hoop, in ieder geval in het genomisch onderzoek. Het concept van reproduceerbaarheid op onafhankelijke datasets is inmiddels behoorlijk goed verankerd in dit onderzoek: het is moeilijk om een mooie bevinding te publiceren zonder onafhankelijke validatie. Daarom moeten we ons wapenen tegen overoptimisme bij het bestuderen van de initiële dataset. En laat statistici daar nu juist goed in zijn, een kwaliteit die we moeten propageren bij andere wetenschappers, wellicht beter dan we tot nog toe doen.

Referenties

1 C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8 (1936), 1–62.

2 R.A. Fisher, Has Mendel's work been rediscovered? *Ann. of Sci.* 1 (1936), 115–137.

3 R. Lister et al., Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells, Nature 471 (2011), 68–73.

4 G. Mendel, Experiments in Plant Hybridization (1865), beschikbaar via www.mendelweb.org/Mendel.html.

5 M.A. van de Wiel, Fictive replicates in epigenomic analysis (2014), doi:10.1038/nature09798.

6 M.A. van de Wiel, 'Het zit in de genen, toch?', inaugurele rede Vrije Universiteit, 23 mei 2014, www.few.vu.nl/~mavdwiel.

7 Citaat van Hal Varian, hoofdeconoom bij Google. Het exacte citaat varieert enigszins in de media.

8 Origineel citaat: "There are three kinds of lies: lies, damned lies, and statistics." Het citaat wordt vaak onterecht toegeschreven aan Mark Twain, maar de echte bron is onduidelijk. en.wikipedia.org/wiki/Lies,_damned_lies,_and_statistics.