de Correspondent

03.11.2015 • Leestijd 9 - 12 minuten

Statistisch onderzoek: overheden, bedrijven en wetenschap draaien erop. Maar een heel groot deel van de statistische verbanden is gebaseerd op toeval - op net zolang proberen tot je het gewenste resultaat hebt. Hoe herken je dit misleidende gegoochel met cijfers?

Onderzoek wijst uit: 48 procent van de economen zijn spugende lama's

Correspondent Economie

Jesse **FREDERIK**

Foto: Getty Images

Op bezoek bij een grote Amerikaanse zakenbank kreeg de vermaarde econoom Campbell Harvey een opmerkelijk resultaat onder ogen. Een analist had een ontdekking gedaan waar iedere beurshandelaar van droomt. Na eindeloos gepiel met een overweldigende hoeveelheid cijfers had hij de heilige graal gevonden: een model dat de beurskoers kon voorspellen.

Campbell was razend nieuwsgierig. Wat was die heilige graal? Eigenlijk was het heel simpel, vertelde de analist hem. De index van de Amerikaanse industriële productie van zeventien maanden eerder vertoonde namelijk een verbluffend verband met de koersen. Als je de beurskoers wilde voorspellen, hoefde je dus alleen maar te kijken hoe de industrie anderhalf jaar geleden presteerde.

‘Zeventien maanden geleden?’ vroeg Harvey verbaasd. ‘Dat is wat ongebruikelijk - één, misschien twee maanden geleden à la - maar waarom zeventien maanden?’

‘Dat was de enige die werkte,’ aldus de onderzoeker.

Wat 'werkte'?

Campbell vertelde dit verhaal niet om beleggers van goed advies te voorzien. Hij vertelde het om te laten zien hoe misleidend economisch onderzoek kan zijn.

De vraag is namelijk: wat bedoelde de analist met ‘het werkte’?

Hij verwees naar een ogenschijnlijk saai begrip in de statistiek - de zogenoemde ‘statistische significantietest’ - die de alfa en omega van talloze wetenschappers is geworden.

‘Statistische significantie’ is de toets der ‘waarheid.' En de gevolgen daarvan zijn immens: er worden bergen onzinnig onderzoek gepubliceerd, investeerders gebruiken beleggingsstrategieën die niet werken en overheden gaan de mist in.

Toegegeven: dit klinkt misschien wat overdreven. Maar dat is het niet. Om dat te begrijpen, zal ik eerst uitleggen wat ‘statistische significantie’ precies is.

Statistische significantie uitgelegd

Het was de briljante wiskundige Ronald Fisher die de zogenoemde ‘significantietest' verzon. Hij kwam op het idee toen een vriendin een kop thee met melk weigerde. De melk was namelijk achteraf aan de thee toegevoegd. Niet te drinken, vond ze, melk moest juist vóór het inschenken van de thee worden toegevoegd. ‘Tuurlijk maakt het geen verschil,’ reageerde Fisher verbaasd. Maar zijn vriendin hield voet bij stuk; ze bezwoer het verschil te kunnen proeven.

Fisher had niet alleen ontdekt dat zijn vriendin het verschil kon proeven tussen de twee kopjes thee, hij had ook een statistische revolutie in gang gezet

En dus was het tijd voor een experiment. Een experiment, met een kieskeurige theedrinker en acht kopjes thee, dat de wetenschap voor eens en altijd zou veranderen.

Kon Fishers vriendin werkelijk het verschil proeven tussen thee met melk die van tevoren was ingeschonken en thee met melk die naderhand was toegevoegd? Fisher begreep dat één experiment met één kopje thee geen uitsluitsel kon bieden. Zelfs een lama die zijn keuze kenbaar zou maken door willekeurig in een kop thee te spugen, zou in 50 procent van de gevallen goed zitten.

Door meerdere koppen thee voor te schotelen, kon Fisher het toeval inperken. Uiteindelijk besloot hij acht koppen thee voor te schotelen, waarvan aan vier van tevoren melk was toegevoegd en vier achteraf. De kans dat zijn vriendin dan bij toeval vier keer het juiste antwoord gaf, zo rekende Fisher uit, was nog maar 1,4 procent. Deze kans noemde Fisher de p-waarde. Het verschil tussen een kundige theeproever en een spugende lama werd bij zo'n lage p-waarde een stuk duidelijker.

Fisher gebruikte zijn natte vinger en besloot dat wetenschappers voortaan een p-waarde (een kans op een toevalstreffer dus) van maximaal 5 procent zouden moeten aanhouden bij het beoordelen van de betrouwbaarheid van hun onderzoek.

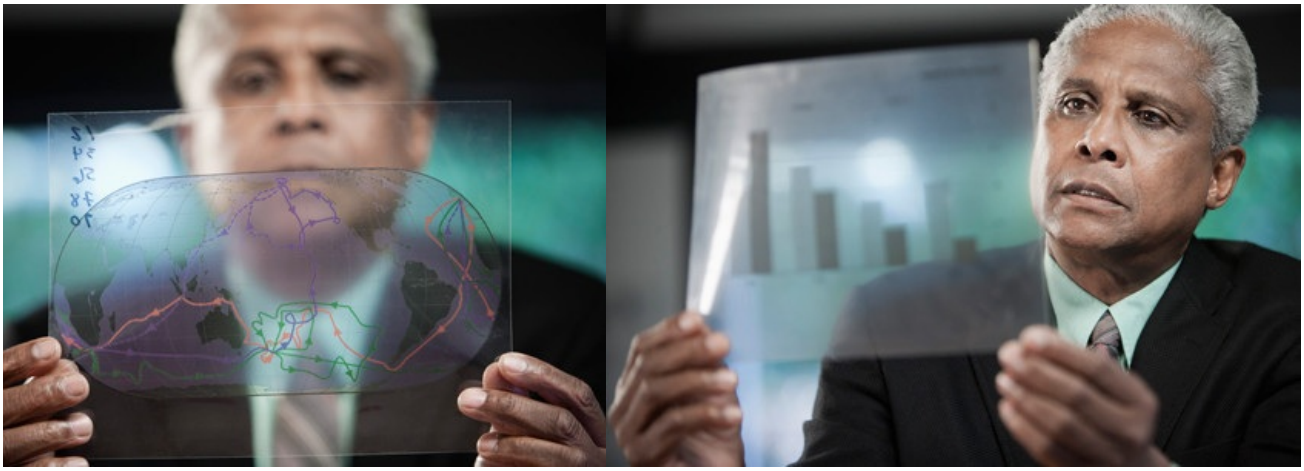
En zo geschiedde. Een p-waarde van minder dan 5 procent is de horde gebleven waar sociale en medische wetenschappers overheen moeten springen. Fisher had niet alleen ontdekt dat zijn vriendin inderdaad het verschil kon proeven tussen de twee kopjes thee, hij had ook een statistische revolutie in gang gezet.

Fishers ongelijk aangetoond

Je zou zeggen: dankzij Fisher weten we dat minstens 95 procent van het statistische onderzoek klopt. Dat klinkt als een prima score, maar de werkelijkheid is helaas een andere. Natuurlijk, 5 procent kans op één toevalstreffer is klein, net zoals het onwaarschijnlijk is dat je zes keer achter elkaar zes dobbelt.

Maar wat nu als je het de hele dag blijft proberen? Of een hele carrière lang? En wat als een hele beroepsgroep van professionele dobbelaars fulltime probeert die zessen te gooien en het alleen vertelt als het lukt?

Dit is min of meer hoe de economische wetenschap werkt. Er wordt heel, heel veel gedobbeld.



Stel: economen testen 1.100 dingen die weleens van invloed zouden kunnen zijn op de beurskoersen: de groei van de omzet, de hoogte van de winst, de beloning van de bestuursvoorzitter, de stand van Saturnus en ga zo maar door. Laten we zeggen dat er bij 100 van de 1.100 in werkelijkheid een verband bestaat. Met een kans van 5 procent op een toevalstreffer, zullen er van de resterende duizend maar liefst vijftig zijn die alsnog een ‘statistisch significant’ verband laten zien. Wie weet vind je opeens een verband tussen de beurskoers en de stand van Saturnus. Of de industriële productie van zeventien maanden geleden.

Maar er is nog iets aan de hand: zoals je toevallig een verband kunt vinden dat niet bestaat, mis je door toeval soms ook een verband dat wél bestaat. Om een echt verband waar te kunnen nemen, moet je steekproef dus groot én gevarieerd genoeg zijn. In de neurologie, zo schatte een groep prominente wetenschappers op basis van honderden onderzoeken, hebben de meeste onderzoeken bijvoorbeeld maar een kans van 21 procent om een bestaand effect waar te nemen.

Stel dat het in de economie veel beter is. En je van de dingen die echt van invloed zijn op de beurskoersen maar in 60 procent een statistisch significant resultaat vindt. Van de honderd dingen die écht verband hebben met de beurskoers, zijn dat er dus zestig. Samen met de eerdere vijftig toevallige resultaten heb je dus in totaal 110 verbanden aangetoond.

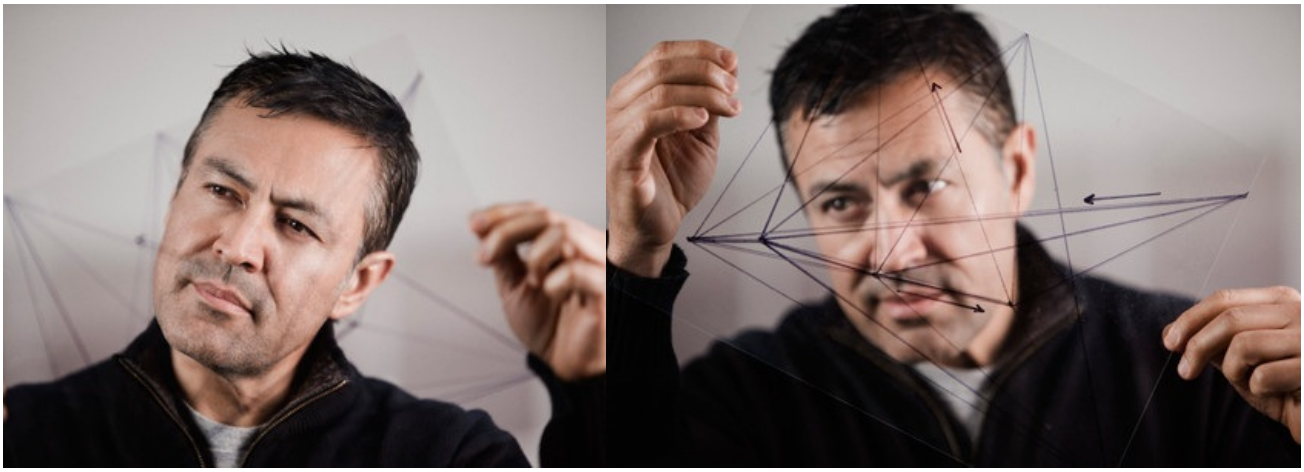
Fisher hoopte dat zijn significantietest ervoor zou zorgen dat 95 procent van de resultaten kloppen. Maar we zien nu dat bijna de helft van die resultaten, 50 van de 110 om precies te zijn, niet klopt.

God dobbelt niet, wetenschappers wel

Maar het wordt nog erger. Meestal weten we namelijk niet eens op hoeveel cijferreeksen de wetenschapper zijn formules heeft losgelaten, hoeveel er is gedobbeld om resultaten te verkrijgen. Dit soort informatie haalt zelden de wetenschappelijke tijdschriften.

Een onderzoek is doorgaans pas publicabel als er een statistisch significant resultaat is. Zo kan het gebeuren dat onderzoekers een heel rijtje afgaan: presteert een bedrijf dat met de letter ‘a’ begint beter op de beurs, nee, de letter ‘b’ dan, nee, dan ‘c,' nee, tot eindelijk ‘d’ wel raak is - eureka! - en een publicatiewaardig resultaat oplevert. (En ja, een onderzoek naar ‘alfabetische vooroordelen’ en de invloed op de beurskoers bestaat echt.)

Het punt is: als je maar lang genoeg zoekt, zul je bij toeval altijd een significant verband vinden. Hier is ook een wetenschappelijke term voor: p-hacken. En dit gebeurt op grote schaal. Niet alleen in dubieuze wetenschappelijke blaadjes, maar ook in de meest prestigieuze toptijdschriften.



De econoom Abel Brodeur bekeek bijvoorbeeld al het onderzoek dat tussen 2005 en 2011 in drie economische toptijdschriften was verschenen. Wat bleek: er waren verdacht weinig p-waardes vlak boven de 5 procent (‘niet-significant’) en verdacht veel p-waardes net onder de 5 procent (‘significant’).

Voor de goede orde: p-hacken is geen fraude. Het gebeurt vaak niet eens bewust. Zo’n 40 procent van de economen gaf onlangs toe te stoppen met onderzoek wanneer het ‘gewenste resultaat’ was gevonden. Zo'n 36,5 procent van de economen gaf aan meer cijferreeksen toe te voegen als het gewenste resultaat nog niet was bereikt. (En dit is dan wat economen zelf toegeven in peilingen - in werkelijkheid is het probleem waarschijnlijk nog groter.)

Dit zal alleen maar meer voor gaan komen. Door razendsnelle computers en de enorme hoeveelheid cijfers die onderzoekers tegenwoordig tot hun beschikking hebben, wordt p-

hacken alleen maar makkelijker. Moesten onderzoekers vroeger hun berekeningen op een telraam doen, tegenwoordig druk je op één knop en heb je een resultaat. Winkelen in de data is gemakkelijker en goedkoper dan ooit.

'De meeste resultaten in de financiële economie kloppen niet'

Neem de beurskoersen: aan het begin van de jaren zeventig waren er slechts tien factoren ontdekt die een verband vertoonden met de koersen. In 2012 waren het er al 314. Maar hoeveel daarvan zijn het gevolg van p-hacken? Het laaghangende fruit (de beurskoers gaat omhoog als de winst omhoog gaat, duh) is al geplukt. Tegenwoordig komen er steeds exotischere verklaringen bij. Op sommige dagen, maar niet altijd dezelfde, blijkt de beurs beter te presteren, bleek bijvoorbeeld uit onderzoek.

De fixatie op de p-waarde heeft, kortom, een hoop onzinnige resultaten opgeleverd. Campbell Harvey komt tot een vernietigende conclusie: ‘De meeste resultaten in de financiële economie, of ze nou gepubliceerd zijn in academische tijdschriften of gebruikt worden als handelsstrategie door een vermogensbeheerder, kloppen niet.’

En dat heeft belangrijke gevolgen. Want pensioenfondsen en verzekeraars betalen miljarden aan fondsbeheerders omdat ze beweren de markt te kunnen verslaan. Maar volgens Harvey zijn de meeste van deze claims aantoonbare onzin. ‘Ongeveer de helft van de financiële producten die een beter resultaat beloven en die [vermogensbeheerders, JF] aan hun klanten verkopen werkt niet.’

Dat geldt voor veel meer economisch onderzoek. Zo beschrijft de econoom Morten Jerven hoe er sinds het begin van de jaren negentig een stortvloed aan verklaringen voor het achterblijven van de Afrikaanse groei is gevonden. In twintig jaar tijd zijn er maar liefst 145 variabelen gevonden - van geografie tot cultuur, van ziekte tot ontwikkelingshulpverslaving, van de mate van eigendomsrecht tot de defensieuitgaven - allemaal zouden ze invloed hebben op de economische groei.

Deze verklaringen worden gebruikt voor beleid, maar we weten niet goed of ze werkelijk kloppen. Ook hier geldt: een hoop verklaringen zijn onvermijdelijk toevalstreffers.

Wat te doen?

Gelukkig zijn er oplossingen.

De eerste ligt het meest voor de hand: stel gewoon een hogere significantie-eis. Niet voor niets gebruikten natuurkundigen in de zoektocht naar het Higgsdeeltje geen significantiegrens van één op twintig, maar van één op drieëneenhalf miljoen. Dat heeft een eenvoudige reden: er zijn in de deeltjesversneller in Genève al zo veel tests uitgevoerd dat je

bij een te lage significantiegrens het Higgsdeeltje aan de lopende band zou vinden.

Een tweede stap kan zijn om wetenschappers minder vrijheid te geven om hun methodes en data te manipuleren. Dit kan gedaan worden door middel van een ‘pre-analyseplan.' De wetenschapper geeft van tevoren aan wat hij van plan is, welke methoden hij wil gebruiken en wat zijn hypothese is. Mocht deze pre-analyse geaccepteerd worden, dan zegt het wetenschappelijk tijdschrift toe het resultaat - wat dat ook moge zijn - te publiceren.

Een derde oplossing is om meer zogenoemde ‘replicaties’ uit te voeren. Dat kan twee dingen betekenen. Pure replicaties: hetzelfde onderzoek nog eens doen met precies dezelfde data en precies dezelfde methoden om te kijken of er geen fouten zijn gemaakt. Of statistische replicatie: de hypothese nog eens testen met nieuwe cijferreeksen en/of andere methodes.

Replicaties kunnen regelmatig bestaande inzichten ondergraven. Een paar jaar geleden repliceerde een jonge wetenschapper bijvoorbeeld een toonaangevende studie van het economenduo Carmen Reinhart en Kenneth Rogoff. De resultaten uit dit onderzoek, die dikwijls werden aangehaald door bezuinigingsgezinde politici, bleken deels gebaseerd op een tikfout.

Replicaties komen echter zelden voor in de economie. Bovendien is repliceren vaak onmogelijk omdat de data en de gebruikte berekeningen niet beschikbaar zijn. Toen twee economen van de Amerikaanse centrale bank onderzoek uit een van de grootste financiële vaktijdschriften over probeerden te doen, bleek dat ze slechts 29 van de 59 artikelen konden repliceren. De oorspronkelijke auteurs wilden maar al te vaak niet meewerken met de replicatie.

Kortom: er is een grote verandering in de economische wetenschap nodig. Een verandering waarbij het moet lonen om te repliceren en om pre-analyseplannen in te dienen. Want leek de industriële productie van zeventien maanden geleden al obscuur - het zal alleen maar erger worden als we niets doen tegen p-hacken. Campbell Harvey voorspelt dat we met het huidige tempo van ‘ontdekking’ tegen 2025 maar liefst 600 variabelen hebben die de beurskoersen verklaren.

En als consument van economisch onderzoek? Wees sceptisch - vooral als het om slechts één onderzoek gaat en het aantal mogelijke verklaringen eindeloos is. Je kunt soms net zo goed op een spugende lama afgaan.

Meer verhalen hierover: