Rob Flohr

# Statistische significantie in wetenschapsfilosofisch perspectief

Stenden Working Papers in Service Studies
SWP 2010/04

# Stenden Working Papers in Service Studies

A working paper summarizes original research in a field of study, and is intended for publication within a period of one to three years. Stenden University of Applied Sciences (www.stenden.com) has initiated an own series of Working Papers in 2008.

The Stenden Research Group in Service Studies publishes a dedicated series of working papers. Aim of the Research Group in Service Studies is to offer a venue for reflection on the role of services in society. Its ultimate aim is to contribute to a better quality of life in our society through superior, sustainable services.

The service sector is very diverse. The Research Group in Service Studies focuses chiefly on hospitality; retailing; tourism; leisure and media. This choice is guided by the fact that Stenden offers education in these particular lines of business. Our enquiry is conducted from different perspectives such as sustainable development; authenticity; leadership and multicultural sensitivity and focuses both on the business world and on the educational work at Stenden University itself.

All Stenden Working Papers have passed through a process of peers' review at one of Stenden Research Groups. Even though, working papers are still in draft form and are distributed for purposes of comment and discussion only. They may not be reproduced or quoted without permission of the copyright holder. All Stenden Working Papers are copyrighted by their authors. For permission to reproduce or to request a copy, contact an author directly.

A SWP Number is a unique identifier assigned to that Stenden Working Paper.

This specific article is written in Dutch.

# STATISTISCHE SIGNIFICANTIE IN WETENSCHAPSFILOSOFISCH PERSPECTIEF

Rob Flohr

Stenden University of Applied Sciences
Rengerslaan 8
8917 DD Leeuwarden – The Netherlands
+31-58-2441441

rob.flohr@stenden.com

---

We gratefully acknowledge the financial support of the Stenden Research Group in Man and Organisation to complete this research.

---

Abstract:

De ontwikkeling van statistische softwarepakketten heeft het onder meer mogelijk gemaakt dat op tamelijk eenvoudige wijze kan worden vastgesteld of een verkregen onderzoeksresultaat statistisch significant is of niet. Het statistische significantiebegrip is echter veel complexer dan doorgaans aangenomen wordt. Dit brengt het gevaar met zich mee dat aan de vaststelling van significantie onjuiste conclusies en incorrecte uitspraken over de werkelijkheid verbonden kunnen worden.
In dit artikel wordt beschreven hoe de invulling van het statistische significantiebegrip in de loop van de tijd tot stand gekomen is, en welke grondgedachten en praktische overwegingen hierbij een rol gespeeld hebben. Het artikel is bedoeld als een pleidooi voor een reflexief gebruik van het significantiebegrip en voor een hernieuwde bezinning op de wijze waarop overwegingen van theoretische aard enerzijds en dataverzameling en -interpretatie anderzijds op elkaar betrokken kunnen worden bij het formuleren van wetenschappelijk verantwoorde uitspraken over de werkelijkheid.

Keywords: significantie, p-waarde, kansbegrip

---

*To understand God's thoughts, we must study statistics,*
*for these are the measure of his purpose.*
Florence Nightingale

## VOORWOORD

Als lid van de kenniskring Mens en Organisatie, onder leiding van lector dr. Adriaan Bekman, doe ik onderzoek naar de verschillende manieren waarop zingeving tot uitdrukking gebracht kan worden. Wanneer we zingeving relateren aan het gegeven dat de mens zich op enigerlei wijze tracht te verhouden tot de wereld waarin hij gesitueerd is, en van daaruit betekenissen verleent aan gebeurtenissen en ervaringen, kunnen we het vak statistiek zien in het licht van de poging van de mens om op een rationele wijze om te gaan met de onzekerheid die inherent is aan zijn bestaan.

---

Statistische methoden en technieken zijn vaak ontwikkeld als antwoord op concrete vraagstukken in de praktijk van alledag. Ze worden vervolgens op grote schaal gebruikt bij het doen van onderzoek van velerlei aard. Statistiek kan dus met recht gezien worden als een vorm van toegepaste wetenschap. Bij toegepaste wetenschap dringt de vraag zich op hoe funderende theoretische concepten en praktische bruikbaarheid zich tot elkaar verhouden. In het geval van statistiek ligt het gevaar op de loer dat onderzoek versmald wordt tot het toepassen van statistische technieken zonder in voldoende mate te reflecteren op onderliggende betekenissen. De ruime beschikbaarheid van statistische software, op zich een enorme verworvenheid voor onderwijs en onderzoek, vergroot dit gevaar. In het verlengde hiervan kan de vraag gesteld worden in hoeverre de nadruk die de toepassing van statistische technieken in het kader van onderzoek momenteel krijgt, ten koste gaat van aandacht voor theorievorming.

De ontwikkeling van statistische software heeft het onder meer mogelijk gemaakt dat onderzoekers, docenten en studenten statistische toetsen kunnen gebruiken in het kader van kwantitatief empirisch onderzoek. Het begrip significantie speelt hierbij een belangrijke rol omdat op basis van dit begrip doorgaans uitspraken over de werkelijkheid geformuleerd worden. In dit artikel wordt het statistische significantiebegrip in een historisch en wetenschapsfilosofisch perspectief geplaatst. Het doel is duidelijk te maken hoe in de loop van de tijd misverstanden en onjuiste interpretaties ten aanzien van het significantiebegrip konden ontstaan en dat het van belang is om daar ook nu nog op bedacht te zijn.
Eerst wordt een korte schets gegeven van het falsificatieprincipe van de filosoof Karl Popper aangezien dit principe de aanzet gegeven heeft tot het ontwikkelen van het huidige significantiebegrip. Vervolgens worden in het verlengde hiervan de denkbeelden uiteengezet van de statistici R.A. Fisher, J. Neyman en E.S. Pearson, die beschouwd kunnen worden als grondleggers van de moderne statistiek. Omdat statistiek ondenkbaar

is zonder het kansbegrip, wordt daarna stilgestaan bij mogelijke benaderingen van het kansbegrip in relatie tot de onzekerheid die inherent is aan het menselijk bestaan. Het artikel sluit af met een beschouwing omtrent de huidige stand van zaken ten aanzien van het significantiebegrip en enkele suggesties ten aanzien van de relatie tussen statistische toetsen en theorievorming bij het doen van onderzoek.

## HET SIGNIFICANTIEBEGRIP IN RELATIE TOT HET FALSIFICATIEPRINCIPE VAN KARL POPPER

De vragen die in een bepaalde periode aan de orde zijn verwijzen doorgaans naar vragen die daaraan voorafgingen en de antwoorden die in dat verband geformuleerd zijn. Voor een goed begrip van een bepaalde filosofische opvatting of theorie zouden we daarom in principe vele eeuwen terug in de tijd moeten gaan. De vraag die ten grondslag ligt aan het onderwerp van dit artikel betreft de bronnen van onze kennis en gaat terug tot Plato (427-347 v.Chr.) en Aristoteles (384-322 v. Chr.) als representanten van respectievelijk het rationalisme en het empirisme. Volgens het rationalisme moet alles wat kennis wil heten afleidbaar zijn uit door de menselijke rede voortgebrachte eerste beginselen terwijl volgens het empirisme alle kennis op ervaring gefundeerd moet zijn. Binnen het bestek van dit artikel neem ik echter het begin van de vorige eeuw als beginpunt.

Vertegenwoordigers van het zogeheten logisch positivisme[1] als Moritz Schlick (1882-1936), Otto Neurath (1882-1945), Rudolf Carnap (1891-1970) e.a. hadden zich ten doel gesteld om de wetenschap te voorzien van een onaantastbaar fundament waarop het bouwwerk van onze kennis opgetrokken zou kunnen worden. Door zich alleen op empirische uitspraken te richten zou de wetenschap vrij gemaakt kunnen worden van speculaties van metafysische aard. De logisch positivisten zochten naar een demarcatiecriterium dat wetenschap en niet-wetenschap van elkaar zou kunnen onderscheiden en vonden dat in het verificatiebegrip: wetenschappelijke uitspraken moeten geverifieerd kunnen worden, bevestigd kunnen worden door de feiten: *"Building on an exclusively empirical basis would rid science of metaphysical speculation."* (Dooremalen e.a. 2007: 176).
Dan verschijnt Popper ten tonele. Karl Raimund Popper (1902-1994), later Sir Karl, deelde de opvatting van de logisch positivisten aangaande het belang van logica en wiskunde voor de wetenschap en hij kon zich ook helemaal vinden in de nadruk die zij legden op het empirisch toetsen van wetenschappelijke uitspraken. Popper formuleerde echter ook een aantal fundamentele bezwaren tegen het logisch positivisme. In dit artikel zal ik me beperken tot één belangrijk element, namelijk zijn falsificatieprincipe.[2] Al in zijn *Logik der Forschung* uit 1935, in 1959 in het Engels verschenen onder de titel *The Logic of Scientific Discovery*, had Popper dit principe geformuleerd dat op het volgende neerkomt: de wetenschap moet zich niet inspannen om steeds weer nieuw bewijsmateriaal te zoeken om een theorie te verifiëren, maar moet daarentegen

[1] Logisch positivisme: wetenschappelijke uitspraken zijn of logisch en wiskundig, of feitelijk van aard.
[2] Voor een eerste kennismaking met de filosofie van Popper, zie Dooremalen e.a. 2007:195-236.

proberen de eigen hypothese te ontkrachten. Alleen daardoor kan de wetenschap een stap verder komen. Absolute zekerheid is onhaalbaar, een mens komt nooit verder dan hypothesen omtrent de werkelijkheid. Het voorlopige karakter van alle kennis maakt dat we stap voor stap te werk moeten gaan. *"Popper avows that the only way to solve practical problems, in the natural as well as the social sciences, in politics as well as daily life, is the method of 'piecemeal engineering': the method of making 'small adjustments and readjustments which can be continually improved upon', changing one aspect at a time on a local scale".* (Dooremalen e.a. 2007: 223). Gesloten wereldbeelden zijn gevaarlijk en dienen bestreden te worden. De open houding van de wetenschapper was voor Popper van cruciaal belang.

De achtergrond van Popper's filosofie wordt voor een belangrijk deel gevormd door zijn kennismaking met het marxistische gedachtegoed (als 16-jarige scholier!) en zijn ontmoetingen met de psycholoog Alfred Adler. Popper kwam tot de conclusie dat zowel het marxisme als de psycho-analyse van Adler met de astrologie gemeen hebben, dat ze alles kunnen verklaren en dat ze altijd door de feiten bevestigd kunnen worden. Dit komt doordat binnen deze theorieën elk feit kan worden geïnterpreteerd als een bevestiging van de theorie. Zo kon Adler met zijn concept van het minderwaardigheidscomplex de meest uiteenlopende gedragingen verklaren. Maar als gevolg daarvan, aldus Popper, zijn de uitspraken die vanuit dergelijke theorieën geformuleerd worden in feite betekenisloos. Bovendien zou een gerespecteerde wetenschap als de natuurkunde op grond van het verificatieprincipe het predikaat wetenschappelijk niet verdienen aangezien geen enkele theorie ooit volledig geverifieerd kan worden. Popper's beroemd geworden voorbeeld van de witte zwanen houdt in dat, hoeveel witte zwanen je tot op heden ook waargenomen hebt, het nooit uitgesloten kan worden dat de volgende zwaan zwart is. Het is logisch onmogelijk om de uitspraak 'alle zwanen zijn wit' te verifiëren.[3] Het moge duidelijk zijn dat het verificatiecriterium van de logisch positivisten geen deugdelijk demarcatiecriterium kan zijn en dat Popper's falsificatiecriterium een alternatief biedt. Er diende zich echter een probleem aan.

## HET FALSIFIEREN VAN STATISTISCHE HYPOTHESEN: EEN PROBLEEM

Het betreft het volgende. Het in beginsel eenvoudige mechanisme waarbij een hypothese, onder bepaalde voorwaarden, op grond van waarnemingen kan worden gefalsifieerd, gaat niet op voor statistische hypothesen. Een hypothese van statistische aard kent namelijk waarschijnlijkheden of kansen toe aan mogelijke gebeurtenissen en bevat geen uitspraak omtrent het al dan niet optreden van één specifieke gebeurtenis. Tegelijkertijd hebben theorieën van statistische aard een belangrijke plaats binnen de wetenschap ingenomen en worden ze regelmatig getoetst. Popper was zich van dit probleem bewust. Zo schrijft hij in *The Logic of Scientific Discovery: "According to my view, however, probability statements, just because they are completely undecidable, are metaphysical unless we decide to make them falsifiable by accepting a methodological rule."* (Popper 1975: 262). Het is van belang op te merken dat een nieuw element hier zijn intrede doet. Blijkbaar spelen volgens Popper bij het empirisch toetsen van

[3] Dit wordt het inductieprobleem genoemd waarbij inductie verwijst naar de redeneerwijze waarin men op grond van een waargenomen regelmaat in een beperkt aantal gevallen overgaat tot een universele uitspraak waarin zo'n regelmaat voor alle dergelijke gevallen wordt gesteld.

statistische hypothesen niet alleen wetenschapsfilosofische overwegingen omtrent de relatie tussen theorie en werkelijkheid een rol, maar ook methodologische regels die ons in staat stellen om die empirische toetsing uit te voeren. Naar mijn mening is deze dualiteit van, wat ik zou willen noemen, wetenschappelijke vragen enerzijds en beslissingsvraagstukken ('decision making problems') anderzijds, of zo u wilt, van epistemologische resp. methodologische vragen, karakteristiek voor het concept van statistische significantie-toetsing waarop ik verder in het artikel uitgebreid in zal gaan, tenminste voor zover het de huidige, klassieke statistiek betreft.[4]

Popper zocht de oplossing in een beslissing van methodologische[5] aard die een *practical falsification* mogelijk zou kunnen maken. *"And a physicist is usually quite well able to decide whether he may for the time being accept some particular probability hypothesis as 'empirically confirmed', or whether he ought to reject it as 'practically falsified', i.e., as useless for purposes of prediction. It is fairly clear that this 'practical falsification' can be obtained only through a methodological decision to regard highly improbable events as ruled out – as prohibited."* (Popper 1975: 191). Met andere woorden, een statistische hypothese is weliswaar strikt genomen niet falsifieerbaar, maar wanneer een gebeurtenis optreedt of een waarneming wordt gedaan die op grond van die hypothese erg onwaarschijnlijk is, zouden we die hypothese moeten verwerpen.
Poppers oplossing was echter niet correct. Het is namelijk juist kenmerkend voor een statistische hypothese dat gebeurtenissen die erg onwaarschijnlijk zijn op grond van die hypothese, niet uitgesloten kunnen worden. Neem het voorbeeld van het kansexperiment waarbij we een munt 10.000 keer opgooien om te toetsen of het een zuivere munt betreft of niet. Als het een zuivere munt is, is de kans op één specifieke uitkomst, in dit geval een specifieke reeks – de vakterm is 'permutatie'- van kop en munt, K en M, in totaal 10.000 letters, heel erg klein: $\left(\frac{1}{2}\right)^{10000}$. Toch is het de kans op elke mogelijke uitkomst waarvan er één ook daadwerkelijk zal plaatsvinden. Een uitkomst waaraan op grond van de te toetsen hypothese een hele kleine kans wordt toegekend, kan dus niet uitgesloten worden. Hoog tijd dus om de volgende hoofdrolspeler in dit verhaal voor het voetlicht te brengen: de eminente statisticus Fisher.

Sir Ronald Aylmer Fisher (1890-1962) wordt algemeen beschouwd als een van de grondleggers van de hedendaagse klassieke of frequentistische (zie voetnoot 5) statistiek.[6] Hij werd geïnspireerd door zowel de falsificatietheorie van Popper als door het ideaal van objectieve wetenschap (Howson & Urbach 2006). Hij ontwikkelde zijn theorie van de significantietoets om theorieën en hypothesen van statistische aard te kunnen

[4] De klassieke of frequentistische (afgeleid van relatieve frequentie, de term wordt verderop in de tekst uitgelegd) statistiek is al ongeveer een eeuw de dominante vorm van statistiek. Een interessant alternatief wordt gevormd door de zogeheten Bayesiaanse statistiek die sterk aan betekenis wint. Het voert in het bestek van dit artikel te ver om op de Bayesiaanse statistiek in te gaan. De geïnteresseerde lezer zij verwezen naar O'Hagan & Luce 2003 of, voor een diepgaander behandeling van de Bayesiaanse statistiek en haar methoden, naar Hoijtink, H., Klugkist, I., Boelen, P.A. (Eds.).(2008). *Bayesian Evaluation of Informative Hypotheses.* New York: Springer en W.M.Bolstad (2007). *Introduction to Bayesian Statistics*. Hoboken, New Jersey:John Wiley. De auteur van dit artikel werkt momenteel aan een Nederlandstalige inleiding tot de Bayesiaanse statistiek.
[5] Methodologie omvat regels over de methode (methodos = de weg waarlangs te gaan) die gezien kunnen worden als spelregels voor het sociale proces dat wetenschapsbeoefening heet, zie P.G. Swanborn (2005). *Methoden van sociaal-wetenschappelijk onderzoek.* Meppel: Boom, p. 29.
[6] Dus niet van de Bayesiaanse statistiek, zie voetnoot 3.

toetsen. In tegenstelling tot Popper werkte Fisher niet met een minimale waarschijnlijkheid om een statistische hypothese te verwerpen, maar stelde hij voor om dit pas te doen wanneer de uitkomst van een kansexperiment (*experimental evidence*) deel uitmaakt van een verzameling van mogelijke uitkomsten die, op basis van de juistheid van de hypothese, relatief onwaarschijnlijk zijn, dat wil zeggen ten opzichte van andere mogelijke uitkomsten van het experiment. Ter verduidelijking nemen we weer het voorbeeld van het werpen van een munt. Stel dat we de hypothese willen toetsen, door Fisher de nulhypothese genoemd, dat het een zuivere munt betreft. Om de toets uit te voeren formuleren we het volgende kansexperiment: we gooien de munt 20 keer en noteren telkens het resultaat. De analyse doorloopt de volgende drie stappen.

1)Bepaal alle mogelijke uitkomsten van het kansexperiment. De verzameling van alle mogelijke uitkomsten wordt de uitkomstenruimte (*sample space* of *outcome space)* genoemd. In ons voorbeeld gaat het om $2^{20} = 1.048.576$ mogelijke reeksen (permutaties) van K(op) en M(unt), 20 in totaal[7]. Om de mogelijke uitkomsten aan te duiden, kunnen we het aantal keren dat we kop gooien gebruiken, variërend van 0 tot 20. Zo'n numerieke aanduiding of beschrijving van een mogelijke uitkomst wordt een toetsingsgrootheid (*test-statistic*) genoemd. In dit voorbeeld is de toetsingsgrootheid dus het aantal keren kop.[8]

2) Bepaal de kans op elke mogelijke uitkomst van het kansexperiment, op basis van de juistheid van de nulhypothese (dus, in ons voorbeeld, uitgaande van een zuivere munt).[9]

3) Kijk naar alle mogelijke uitkomsten van het kansexperiment – nog steeds op basis van de juistheid van de nulhypothese (dit is een essentieel element!) – die een kans hebben die gelijk is aan of kleiner is dan de kans op de uitkomst die we in het experiment hebben aangetroffen.
Toelichting: stel dat het experiment van 20 keer gooien met de munt een reeks heeft opgeleverd met 4 keer kop en 16 keer munt. De kans op 4 keer kop is gelijk aan 0.0046. De mogelijke uitkomsten met een kans kleiner dan of gelijk aan de kans op de feitelijk gerealiseerde uitkomst van 4 keer kop zijn:
4, 3, 2, 1, 0 en 16, 17, 18, 19 en 20 keer kop (deze uitkomsten zijn even extreem of nog extremer dan 4 keer kop). De kans dat één van deze gebeurtenissen optreedt is gelijk aan de som van de kansen op elk van deze gebeurtenissen en is gelijk aan 0.012. Deze laatste kans wordt de overschrijdingskans of p-waarde (*p-value*) genoemd.

Fisher had het concept van de p-waarde ontwikkeld om van data, verkregen op basis van één enkele steekproef of experiment, te kunnen zeggen in welke mate ze tegen de nulhypothese pleiten, ofwel de nulhypothese kunnen falsifiëren. Zijn doel was

[7] In het kader van dit artikel, waarin de grondgedachten rond statistische significantietoetsen centraal staan, laat ik de berekeningen achterwege. De geïnteresseerde lezer zij verwezen naar een willekeurig leerboek statistiek of kan de auteur per email om de berekeningen vragen (rob.flohr@stenden.com).
[8] Zo'n toetsingsgrootheid wordt een kansvariabele, ook wel stochastische variabele of kortweg stochast genoemd (*random variable)*, gedefinieerd als de numerieke beschrijving van de uitkomst van een kansexperiment.
[9] Voor de berekening van de kansen van de mogelijke uitkomsten van het kansexperiment hebben we een specifieke kansverdeling nodig. Wat voor kansverdeling we nodig hebben hangt af van het type kansexperiment. In ons voorbeeld gebruiken we de zogeheten binomiale kansverdeling die echter onder bepaalde voorwaarden door de normale verdeling benaderd kan worden. Er bestaan veel verschillende kansverdelingen.

epistemisch van aard, dat wil zeggen gericht op kennis omtrent de werkelijkheid: op basis van de gevonden steekproefuitkomst kan een uitspraak over een 'stand van zaken' in de werkelijkheid geformuleerd worden. Hij was op zoek naar een *measure of inductive evidence.*
Zoals we gezien hebben is de p-waarde de kans op de verkregen data, of een nog extremere uitkomst, onder de veronderstelling dat de nulhypothese waar is en dat het toeval de kans op elke mogelijke uitkomst bepaalt op grond van een specifieke kansverdeling. Nu bestaan er nogal wat misverstanden omtrent de precieze betekenis van de p-waarde. Een veel voorkomende maar incorrecte interpretatie luidt dat de p-waarde de kans is dat de verkregen steekproefdata het gevolg van toeval zijn. De p-waarde zelf is immers al berekend onder de aanname dat het toeval alle mogelijke uitkomsten bepaalt.
Wat je wel kunt zeggen is dat je, op basis van de gevonden p-waarde, op rationele gronden een beslissing kunt nemen omtrent het al dan niet accepteren van het idee dat de verkregen data het gevolg zijn van toeval. De onzekerheid die daaraan verbonden is kun je echter niet in een kans uitdrukken (Carver 1978).

Het concept van de p-waarde levert een aantal problemen op.
Het betreft onder meer de omstandigheid dat de toepassing van de wiskunde (lees kansrekening), je zou het ook de wiskundige modellering van onzekerheid kunnen noemen, alleen onder zeer speciale omstandigheden tot exacte uitkomsten leidt. Een belangrijke voorwaarde betreft het aselecte karakter van de steekproef (*randomness*). In de werkelijkheid van de empirische wetenschappen is het aselecte karakter vrijwel onmogelijk te realiseren. Nu hoeft dat op zich nog geen ramp te zijn, in veel gevallen kan met een goede benadering worden volstaan. Maar het is ook mogelijk dat de juistheid van de getrokken conclusies op de tocht staat.
Een ander punt betreft het karakter van de populatie. De meeste wiskundig georiënteerde statistici definiëren statistische vraagstukken in termen van een herhaalde steekproeftrekking uit *één en dezelfde* populatie.[10] Dit maakt de toepassing van een vrij eenvoudig wiskundig model mogelijk, maar het beantwoordt echter niet aan de behoeften van de onderzoeker omdat je over het algemeen niet te maken hebt met één en dezelfde populatie. Stel dat je twee antibiotica vergelijkt ten aanzien van hun werking bij het behandelen van een infectie. Stel dat de significantietoets uitwijst dat het ene antibioticum beter is dan het andere. In welke mate helpt dat ons verder wanneer de toets bijvoorbeeld is uitgevoerd in Amsterdam in 2010 en je wilt het gevonden resultaat benutten voor het gebruik in Afrika in 2011? Misschien is het voorbeeld enigszins extreem maar het neemt niet weg dat we bij elke toepassing van statistiek moeten inschatten in hoeverre we resultaten, verkregen op tijdstip 1, onder een geheel van omstandigheden A, kunnen gebruiken als leidraad voor wat er zal gebeuren op tijdstip 2, onder een geheel van omstandigheden B. (Fisher zelf benoemde dit als *"sampling from a hypothetical infinite population"* (Kerridge & Kerridge 1998:7).
Samengevat: de statistische theorie, zoals geformuleerd in leerboeken, analyseert wat er kan gebeuren indien we herhaalde, aselecte steekproeven zouden trekken uit één en dezelfde populatie. In de meeste gevallen echter, willen we niet alleen iets zeggen over het verleden, maar ook over de toekomst. En om dat te kunnen doen heb je minstens zo

[10] Voor de lezer die enigszins vertrouwd is met statistiek, het betreft hier de steekproevenverdeling (*sampling distribution*).

veel theoretische kennis van het onderwerp in kwestie nodig als kennis van de kansrekening.

Nu is het zeker niet zo dat we op grond hiervan de statistiek maar vaarwel moeten zeggen. Fisher zou in het genoemde voorbeeld van de twee antibiotica voorgesteld hebben de statistische theorie wel degelijk toe te passen maar benadrukte anderzijds dat het hier niet meer een louter statistisch vraagstuk betreft maar, zoals hij het noemde, een *scientific problem*. Hij maakte een onderscheid tussen *scientific problems* en *decision-making problems*. Een voorbeeld van de laatste soort problemen is volgens hem het verwerpen of accepteren van een hypothese. Wetenschappelijke problemen echter, hebben betrekking op het vinden van een verklaring voor verschijnselen die zijn waargenomen waarbij er doorgaans geen sprake is van het bestaan van één en dezelfde populatie.
Wat betekent dit alles nu voor de onderzoekspraktijk? Welnu, Fisher betoogde dat je nooit mag bouwen op het resultaat van één steekproef of experiment of significantietoets. Fisher was van mening dat je de toets onder zoveel mogelijk verschillende omstandigheden zou moeten herhalen. Als de resultaten onder verschillende omstandigheden consistent zijn, dan kun je erop vertrouwen. Als ze uiteenlopen, denk dan nog eens goed na.

Na Fisher ontwikkelden de statistici Jerzy Neyman (1894-1981) en Egon Sharpe Pearson (1895-1980)- zoon van de bekende statisticus Karl Pearson – hun concept van de hypothesetoetsing. Ze stelden voor om in plaats van een afzonderlijke hypothese te toetsen, de significantietoets te herformuleren in termen van twee (of meer) rivaliserende hypothesen, de nulhypothese en de alternatieve hypothese(n), doorgaans aangeduid met $H_0$ resp. $H_1$ of $H_a$. Hun doel was niet van epistemische maar van gedragsmatige aard; het concept van de hypothesetoetsing is niet gericht op het mogelijk maken van een uitspraak over een stand van zaken in de werkelijkheid, maar op het nemen van een beslissing: het verwerpen of het accepteren van de nulhypothese. Bovendien gingen ze niet uit van één enkele steekproef maar van het idee van een veelvuldige herhaling van de steekproeftrekking waarbij de kans op een foutieve beslissing op lange termijn geminimaliseerd zou worden *(repetitive error rate)*.
De Neyman-Pearson methode is gemeengoed geworden. Hierbij zijn er twee mogelijkheden: een hypothese wordt verworpen of geaccepteerd. Dit houdt in dat er twee soorten fouten gemaakt kunnen worden: $H_0$ wordt verworpen terwijl die in werkelijkheid waar is (fout van de eerste soort) of $H_0$ wordt geaccepteerd terwijl in werkelijkheid $H_1$ waar is (fout van de tweede soort). De Neyman-Pearson methode is erop gericht de kans op beide soorten fouten zo klein mogelijk te houden.[11]
Fisher's concept van de p-waarde enerzijds (ontwikkeld in de 20-er jaren van de vorige eeuw) en de Neyman-Pearson methode van hypothese toetsing anderzijds (ontwikkeld in

[11] De kans op een fout van de eerste soort is gelijk aan het significantieniveau $\alpha$. Om de kans op de fout van de tweede soort,meestal aangeduid met $\beta$, te kunnen berekenen moet je een specifieke waarde voor $H_1$ kiezen. Ik zal in het bestek van dit artikel hier niet verder op ingaan. Wel zij nog vermeld dat het complement van $\beta$, dus 1 - $\beta$, het onderscheidingsvermogen (*power*) van de significantietoets genoemd wordt, zijnde de kans dat de alternatieve hypothese aanvaard wordt wanneer die in werkelijkheid ook juist is.

begin jaren 30 van de vorige eeuw) zijn daarmee twee verschillende benaderingen van het significantiebegrip (Wagenmakers et al 2008). De p-waarde is een maat om de *strength of evidence* te meten terwijl de hypothesetoetsing een methode is om te kunnen kiezen tussen twee hypothesen. Het zijn twee onverenigbare methoden die echter vaak abusievelijk beschouwd worden als elementen van één, coherente benadering van statistisch redeneren (Goodman 1999).
In de loop van de tijd zijn ze met elkaar vervlochten geraakt doordat de p-waarde van Fisher een centrale rol speelt bij het toetsen van de nulhypothese op basis van de Neyman-Pearson methode.
Het is namelijk gebruikelijk, zeker binnen de sociale wetenschappen, om de nulhypothese te verwerpen indien de p-waarde kleiner is dan of gelijk aan 0.05 (hoewel een kritische waarschijnlijkheid van 0.01 of zelfs 0.001 ook kan voorkomen, afhankelijk van het betreffende wetenschapsgebied). Deze kritische waarschijnlijkheid of kans wordt het significantieniveau (*significance level*) van de toets genoemd en wordt aangeduid door de Griekse letter $\alpha$. Wanneer de uitkomst van het kansexperiment zodanig is dat de p-waarde kleiner is dan of gelijk aan $\alpha$ ( $p \leq \alpha$ ), zeggen we dat deze uitkomst significant is op het $\alpha$ significantieniveau en dat dienovereenkomstig de nulhypothese verworpen wordt op het $\alpha$ niveau. In ons voorbeeld van het werpen van de munt wordt de nulhypothese die inhoudt dat het een zuivere munt betreft dus verworpen op het 0.05 of 5% niveau aangezien 0.012<0.05. (Wanneer we 6 keer kop hadden gegooid zou de p-waarde gelijk zijn aan 0.115 en zou de nulhypothese dus niet verworpen worden, althans niet op het 0.05 niveau). (Howson & Urbach 2006: 135).[12]

Als gevolg van de gesignaleerde vervlechting zijn de verschillen tussen het concept van de p-waarde en het concept van de hypothesetoetsing naar de achtergrond verschoven, zo niet geheel verdwenen. Als gevolg daarvan worden vaak verkeerde interpretaties van een verkregen p-waarde geformuleerd. Het vergelijken van de gevonden p-waarde met een gegeven $\alpha$ suggereert een exactheid met betrekking tot de p-waarde die er niet is. Anders gezegd, de p-waarde wordt tegelijkertijd als een *repetitive error rate* en een *measure of evidence* opgevat. Dit laatste leidt dan tot incorrecte kansuitspraken met betrekking tot een of andere 'stand van zaken' in de werkelijkheid, zoals bijvoorbeeld: 'de kans dat de nulhypothese waar is, is 5%', of 'de kans dat de alternatieve hypothese waar is, is 95%', of 'de kans dat de verkregen steekproefuitkomst op toeval berust is 5%' (Carver 1978). Een mogelijke verklaring voor het ontstaan van dit soort incorrecte interpretaties van de p-waarde kan gelegen zijn in de omstandigheid dat de p-waarde en het significantieniveau $\alpha$ op het eerste gezicht veel gemeen hebben: ze kunnen beide in de staart van een kansverdeling liggen en zijn doorgaans van dezelfde orde van grootte. De p-waarde en het significantieniveau $\alpha$ zijn echter wezenlijk verschillend van elkaar: de $\alpha$ wordt, *voorafgaand* aan de dataverzameling, *vastgesteld* op grond van overwegingen van conventionele aard, terwijl de p-waarde pas *achteraf*, nadat de data

[12] Ik heb dit voorbeeld gekozen om het concept van de significantietoets zo duidelijk mogelijk naar voren te brengen. In de dagelijkse praktijk van het statistisch onderzoek gaat het uiteraard niet om kansexperimenten als het gooien van een munt of een dobbelsteen maar om bijvoorbeeld het toetsen van het effect van een medicijn of van een interventie, of om het toetsen van een verschil tussen twee groepen e.d. Hoewel het principe hetzelfde is zoals hierboven geschetst, gaat het dan om andere kansverdelingen. Het kan de normale verdeling betreffen, of bijvoorbeeld de Chi-kwadraat verdeling, of de F-verdeling. In veel gevallen gaat het echter om de t-verdeling, ook wel Student verdeling genaamd, aangezien deze kansverdeling is ontwikkeld door W.S. Gossett die publiceerde onder het pseudoniem ' Student'.

zijn verzameld, wordt *berekend* op basis van een specifieke kansverdeling. De p-waarde van Fisher verwijst naar de vraag: hoe waarschijnlijk is het dat de nulhypothese waar is, gezien mijn steekproefuitkomst? De Neyman-Pearson methode van hypothesetoetsing daarentegen, heeft betrekking op de vraag: is de p-waarde kleiner dan of gelijk aan $\alpha$ of niet? Zo ja, verwerp dan de nulhypothese. Zo nee, accepteer dan de nulhypothese.

De reden dat ik zo uitvoerig ben ingegaan op het concept van de significantietoets is dat een goed begrip van deze toets naar mijn mening noodzakelijk is om de precieze betekenis daarvan te begrijpen zodat we het ook op waarde kunnen schatten. Bovendien hoop ik zo ook duidelijk gemaakt te hebben dat wat we in de praktijk van alledag gebruiken aan statistische toetsen, doorgaans met behulp van SPSS, EXCEL, MINITAB of andere software, in de grond van de zaak het resultaat is van een mengelmoes van verschillende soorten overwegingen. Het is een groot goed dat we heden ten dage over software kunnen beschikken maar het vraagt wel van ons dat we daar op een bedachtzame manier mee omgaan.
Wat betreft die verschillende soorten overwegingen zien we dat, naast overwegingen van (wetenschaps)filosofische aard (zoals het falsificatieprincipe en het ideaal van objectiviteit), ook overwegingen van methodologische (*decision-making)* aard (bijvoorbeeld de ' *practical falsifiability'* van Popper) en van conventionele aard (het bepalen van het significantieniveau) een rol spelen in het proces van statistische significantie-toetsing.[13] Daarnaast hebben we in ons voorbeeld gebruik gemaakt van kansberekeningen. En dat is een verhaal apart. Want wat is een kans eigenlijk? En welke rol kan het kansbegrip spelen in het omgaan met de onzekerheid die inherent is aan ons bestaan? Deze vragen zullen verderop in dit artikel nog aan de orde komen.

We hebben gezien dat een significantietoets uitgevoerd wordt onder de aanname dat de nulhypothese waar is. Dit betekent dat we op basis van een significantietoets niet kunnen beweren dat de nulhypothese waar of onwaar is. Of deze waar of onwaar is, weten we niet en zullen we ook nooit weten. Je zou kunnen zeggen dat begrippen als waarheid en onwaarheid buiten het bereik van het statistische redeneren liggen. We kunnen zelfs niet beweren dat de kans bijvoorbeeld 95% is dat de nulhypothese waar is (of onwaar). Wat kunnen we dan wel beweren? Wat voor soort uitspraak over de werkelijkheid kunnen we op verantwoorde wijze formuleren? Dat valt pas goed te begrijpen wanneer we het kansbegrip, dat aan de significantietoets ten grondslag ligt, uitvoerig uit de doeken hebben gedaan. Vandaar dat ik aan het kansbegrip een aparte paragraaf zal wijden. Voordat ik dat doe, wil ik echter eerst ingaan op de vraag in welke mate de significantietoets ons behulpzaam kan zijn bij ons omgaan met de onzekerheid die inherent is aan ons bestaan. Welke vormen van onzekerheid zijn er en voor welke vorm(en) biedt het significantiebegrip soelaas? En voor welke vorm(en) niet?

## SOORTEN ONZEKERHEID

[13] Dat een $\alpha$ van 0.05 op conventie berust, wil uiteraard nog niet zeggen dat het percentage willekeurig gekozen is. Het heeft te maken met een afstand van twee maal de standaarddeviatie tot het gemiddelde van een verdeling. In het geval van een normale verdeling correspondeert dit, bij tweezijdige toetsing, met een percentage van 4,56%. Het vermoeden bestaat dat Fisher dit heeft afgerond tot 5% (Cowles & Davis 1982).

De Amerikaanse ingenieur en statisticus W. Edwards Deming (1900-1993) maakte een onderscheid tussen e*numerative statistical studies* en *analytic statistical studies* (Deming 1975). Net als Fisher was hij geïnteresseerd in de vraag hoe statistiek en werkelijkheid zich tot elkaar verhouden. Bij het eerste type studies betreft het vraagstukken die heel goed met behulp van de statistische theorie zijn op te lossen, dit soort statistische studies is primair gericht op de beoordeling van resultaten (bijvoorbeeld: heeft een sociale interventie een positief resultaat opgeleverd?). Bij het tweede type studies (de overgrote meerderheid) is een ander soort kennis vereist is, zoals kennis over het onderwerp in kwestie die gebaseerd kan zijn op de jarenlange ervaring van een expert, of kennis gebaseerd op theoretische inzichten, of misschien zelfs professionele intuïtie. Dit soort studies zijn namelijk primair gericht op de verbetering van het proces of van het systeem dat de resultaten in kwestie heeft opgeleverd (de onderliggende vraag is bijvoorbeeld: wat kunnen we zeggen over de kwaliteit van onze sociale interventies?). Deming geeft een eenvoudig criterium om te bepalen of er in een specifieke situatie sprake is van een *enumerative* dan wel *analytic* statistische studie: wanneer de steekproef de hele populatie zou omvatten, dan zou in het geval van een *enumerative study* de onderzoeksvraag volledig beantwoord zijn, terwijl dat in het geval van een *analytic study* niet het geval zou zijn. (Deming 1975:147).
Ingeval van een *analytic study* zijn er twee verschillende vormen van onzekerheid:
- onzekerheid als gevolg van het feit dat we een steekproef onderzoeken en niet de hele populatie. Deze vorm van onzekerheid kunnen we op basis van statistische theorie heel goed kwantitatief uitdrukken.
-onzekerheid als gevolg van het feit dat we uitspraken willen doen over een bepaald tijdstip in de toekomst betreffende een groep (bijvoorbeeld mensen zoals patiënten, werknemers, e.d.) die verschilt van de groep die onze steekproef heeft gevormd. De mate waarin deze vorm van onzekerheid aan de orde is, zal verschillen van vakgebied tot vakgebied. Ze zal waarschijnlijk een kleinere rol spelen bij de natuurwetenschappen dan bij de geneeskunde, sociale wetenschappen en managementwetenschappen aangezien binnen deze laatste vakgebieden veranderingen zich in een relatief hoog tempo voltrekken.

Terugblikkend op het onderscheid dat Fisher maakte tussen *scientific problems* en *decision-making problems* en in het licht van het onderscheid dat Deming maakte, kunnen we mijns inziens concluderen dat de reflectie op de verhouding tussen statistische theorie enerzijds en werkelijkheid anderzijds de nodige aandacht van de onderzoeker verdient. Tevens zou je kunnen stellen dat de vraag hoe statistiek en werkelijkheid zich tot elkaar verhouden goeddeels uit het zicht is verdwenen. Het gevolg daarvan lijkt te zijn dat de tweede vorm van onzekerheid, de onzekerheid die niet wiskundig gearticuleerd kan worden, is geëlimineerd uit het statistische denkkader. Wat daaruit resulteert is gemakkelijker te onderwijzen, maar het gaat voorbij aan de echte vraagstukken waarmee Fisher worstelde en die hij, hoe onvolkomen ook, probeerde op te lossen. Het gevaar ligt op de loer dat we, mede als gevolg van de enorme ontwikkeling van statistische software, zodanig gericht zijn op de toepassing van de statistische theorie, dat we onvoldoende stilstaan bij de vraag welke uitspraken over de werkelijkheid wel en welke niet verantwoord zijn.

## HET KANSBEGRIP

Het denken in termen van waarschijnlijkheden is van betrekkelijk recente datum. Terwijl de negentiende eeuw nog als 'deterministisch' wordt gekarakteriseerd omdat variaties in waarnemingsresultaten werden toegeschreven aan onvolkomenheden in apparatuur e.d., begon in de vorige eeuw het inzicht door te breken dat deze variaties ofwel toevalsuitkomsten, *randomness,* en de daarmee gepaard gaande onzekerheid, inherent zijn aan de werkelijkheid (Salsburg 2002). Darwin's vondst van de biologische variatie heeft hierbij waarschijnlijk een rol gespeeld aangezien toevalsmutaties daarin een centrale rol vervullen.[14]

We dienen een onderscheid te maken tussen de geschiedenis van het kansbegrip en de kansrekening, als onderdeel van de wiskunde, enerzijds en de geschiedenis van de statistiek anderzijds. Wat het kansbegrip betreft hebben opgravingen in het Midden-Oosten en India aangetoond dat al rond 1400 v. Chr. dobbelstenen werden gebruikt. Het duurt echter tot de 16e eeuw voordat het dobbelspel aan een wiskundige analyse onderworpen wordt door de Italiaanse wiskundige Gerolamo Cardano (1501-1576) die, zelf een hartstochtelijk gokker, een handboek voor gokkers schreef.[15] Aan het eind van de 17e eeuw legt de Nederlandse astronoom Christiaan Huygens (1625-1695) de basis voor de huidige kanstheorie en in de zomer van 1654 formuleren Pierre de Fermat en Blaise Pascal de basisbeginselen van de kansrekening in hun briefwisseling naar aanleiding van een vraag van de beroepsgokker Chevalier de Méré.[16] Tijms geeft een fraaie formulering van het belang van kansspelen voor de ontwikkeling van de kansrekening:
"*It is difficult to say who had a greater impact on the mobility of goods in the preindustrial economy: the inventor of the wheel or the crafter of the first pair of dice. One thing, however, is certain: the genius that designed the first random-number generator, like the inventor of the wheel, will very likely remain anonymous forever."* (Tijms 2007: 1).
Hacking situeert de start van het gebruik van het huidige kansbegrip rond het midden van de 17e eeuw omdat dit begrip vanaf die tijd voortdurend opduikt in verhandelingen over praktische en wetenschappelijke vraagstukken. Hij spreekt over *the emergence of probability* en ziet dit als *"…the spontaneous emergence of a new style of thinking about man and God, of describing nations, of existential theology, of merchant adventurers and the methods of science. It has to do with a vast range of new practices of trade and the idea of the state, …, and also with the new sciences of the seventeenth century, with their novel modes of inquiry."* (Hacking 2006, Introduction 2006: VII). Daarvoor werd het begrip kans ook al gebruikt maar toen verwees het vooral naar een mening over de werkelijkheid die in overeenstemming was met de visie van de (kerkelijke) autoriteiten, het begrip verwees dus naar wat door de autoriteiten goedgekeurd kon worden en niet

[14] *"Darwin's theories of evolution postulated that life forms change in response to environmental stress. He proposed that changing environments gave a slight advantage to those random changes that fit better into the new environment. Gradually, as the environment changed and life forms continued to have random mutations, a new species would emerge that was better fit to live and procreate in the new environment. This idea was given the shorthand designation 'survival of the fittest'."* (Salsburg 2002: 18).
[15] Het geloof in een almachtige God die alles bepaalt, bood weinig ruimte voor onderzoek naar toevallige gebeurtenissen (Cowles & Davis 1982).
[16] Zie voor deze briefwisseling het mooie boek van Keith Devlin (2008) *The Unfinished Game. Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern.* New York: Basic Books. De vraag van Chevalier de Méré luidde: kunt u verklaren waarom de weddenschap om met twee dobbelstenen in 24 worpen tenminste één keer dubbel zes te gooien, nadeliger uitvalt dan de weddenschap om met één dobbelsteen in 4 worpen tenminste één keer zes te werpen?

naar "*evidence provided by things".* (Hacking 2006: 32).

Wat betreft de geschiedenis van de statistiek gaat het in Nederland aanvankelijk vooral om de vraag hoe bij het oplossen van praktische vraagstukken van staatkundige aard omgegaan kan worden met grote hoeveelheden gegevens. In Nederland kreeg het woord 'statistiek' ongeveer in het midden van de periode 1750-1850 een plaats in het spraakgebruik en omvatte het onder meer een academische studie van *Staatenkunde* en gekwantificeerde beschrijvingen van naties en regio's . Gedurende lange tijd had statistiek betrekking op niet meer dan eenvoudige metingen in de zin van het systematiseren en kwantificeren van waarnemingen. Gottfried Achenwall (1719-1772), hoogleraar aan de universiteit van Göttingen had *Statistik* of *Staatenkunde* ontwikkeld tot een methode om opmerkelijke feiten betreffende de staat te structureren. Informatie over de politieke kracht van een natie en de werkelijke situatie van het land waren belangrijke doeleinden. Het zal daarom weinig verbazing wekken dat het Italiaanse woord *statista* staatsman betekent (Klep 2002).

In een internationale context moeten we bijvoorbeeld denken aan het combineren van waarnemingen in de sterrenkunde zoals Jacques Cassini in 1740 deed met betrekking tot de hellingshoek van de evenaar ten opzichte van de zon. Naast de sterrenkunde is ook de geodesie een vakgebied waar de statistiek al vroeg een toepassing vindt. Het opstellen van sterftetafels vond al eerder plaats, in de 17e eeuw. Aan het eind van de 19e eeuw komen ook de biologie en de sociale wetenschappen binnen de invloedssfeer van de statistiek (Stigler 1986).
Al met al kun je zeggen dat de opkomst van de statistiek als vak in het teken staat van het ontluiken van een *particular empirical mind* (Klep 2002: 65). Maar pas in de 20e eeuw wordt statistiek een zelfstandig vak waarin het kwantitatief uitdrukken van onzekerheid de centrale plaats inneemt: *"Modern statistics provides a quantitative technology for empirical science; it is a logic and methodology for the measurement of uncertainty and for an examination of the consequences of that uncertainty in the planning and interpretation of experimentation and observation. Statistics, as we now understand the term, has come to be recognized as a separate field only in the twentieth century."* (Stigler 1986: 1). Bij dat tot uitdrukking brengen van het element van onzekerheid speelt het kansbegrip een cruciale rol, vandaar dat we ons nu op dat begrip zullen richten.

Het lastige met het kansbegrip is dat het wiskundig duidelijk gedefinieerd kan worden maar dat het niet mogelijk blijkt om het te definiëren in termen van concrete aspecten van de wereld om ons heen. Eén poging om het begrip kans te definiëren in realistische termen, is door het te plaatsen in de context van de uitkomsten van een experiment dat vele malen herhaald wordt, het begrip kans wordt dan opgevat als een relatieve frequentie.[17] Dit impliceert dat het kansbegrip betekenisloos is wanneer het een

[17] In theorie gaat het om een oneindig aantal herhalingen van een kansexperiment. Wiskundig kan dat uitgedrukt worden met behulp van het limietbegrip. Wanneer $n$ het aantal malen weergeeft dat een kansexperiment herhaald wordt en $f_n$ het aantal keren – de frequentie – dat een bepaalde uitkomst van het kansexperiment gevonden wordt bij $n$ herhalingen van het experiment, dan is de kans in op die uitkomst, in

gebeurtenis betreft die maar één keer kan optreden. Zo kun je de kans van 1/6 op het gooien van een zes met een zuivere dobbelsteen bij benadering opvatten als het aantal worpen dat een zes heeft opgeleverd, gedeeld door het totaal aantal worpen. Het probleem is dat dit alleen mogelijk is wanneer je veronderstelt dat de kans op bepaalde uitkomsten, bijvoorbeeld 1000 keer een zes gooien op een totaal van 1000 worpen, zo goed als nul is. Je veronderstelt daarmee al een kansbegrip.
De huidige statistiek, ook wel klassieke statistiek genoemd[18] zoals ontwikkeld door Ronald Fisher en later door Neyman and Pearson, is gebaseerd op deze frequentistische benadering van het kansbegrip. De implicaties hiervan voor een correcte interpretatie van het significantiebegrip zullen in de volgende paragraaf aan de orde komen.
Een andere poging om het begrip kans in realistische termen te definiëren betreft het klassieke kansbegrip van Pierre-Simon Laplace (1749-1827): het aantal relevante uitkomsten gedeeld door het totaal aantal mogelijke uitkomsten, waarbij deze laatste alle even waarschijnlijk moeten zijn. Maar om te bepalen of uitkomsten even waarschijnlijk zijn zul je in bepaalde gevallen experimenten moeten uitvoeren en daarbij ontkom je niet aan het gebruik van de definitie van kans als relatieve frequentie. Ik zal binnen het bestek van dit artikel niet alle mogelijke benaderingen van het kansbegrip bespreken (zie hiervoor bijvoorbeeld Mellor (2002) en Gillies (2000)) maar ik noem nog wel de subjectieve kans: kans als een *personal degree of belief*.

Kortom, zo concludeert de Amsterdamse hoogleraar wiskunde Ronald Meester, kansen zijn niet te definiëren in realistische termen, de wetenschap maakt weliswaar gebruik van kansen maar we moeten bekennen dat we niet weten wat een kans eigenlijk is (Meester 2004: 34). Hetzelfde geldt volgens hem voor het begrip zwaartekracht. Door het aannemen van het bestaan van de zwaartekracht kunnen we zeer veel verklaren maar we weten niet of de zwaartekracht iets reëels is of alleen een begrip waarmee bepaalde aspecten van de werkelijkheid kunnen begrijpen.

Mellor (2005) onderscheidt drie soorten waarschijnlijkheid en duidt slechts een daarvan aan als kans. Het betreft:
1. *Physical probability*: het begrip 'waarschijnlijk' heeft betrekking op de wereld zelf, op feitelijke zaken. De betreffende uitspraken kunnen ook feitelijk gecontroleerd worden. Voorbeelden hiervan zijn:
-Radiumatomen hebben een halfwaardetijd van ongeveer 1600 jaar, dat wil zeggen dat ze een kans van 50% hebben om binnen die periode te vervallen, en:
-Bij deze munt is het even waarschijnlijk dat je kop als dat je munt gooit.
Deze vorm van waarschijnlijkheid wordt door de eerder genoemde filosoof Carnap de statistische interpretatie van *probability* genoemd en wordt door Gillies (2000) als *objective probability* aangeduid. Het woord objectief verwijst daarbij naar het feit dat ze onafhankelijk is van hoe mensen erover denken. De kans met betrekking tot de radiumatomen bijvoorbeeld bestond al lang voordat er mensen bestonden.
*2. Epistemic probability*: de waarschijnlijkheid verwijst hier niet direct naar iets in de wereld zelf, maar naar uitspraken over de wereld. Dit begrip 'waarschijnlijk' is

---

frequentistische zin, gelijk aan $\lim_{n\to\infty} \frac{f_n}{n}$ . Maar een oneindige herhaling van het experiment is geen realistisch begrip, geen aspect van de concrete werkelijkheid.
[18] Dus niet de Bayesiaanse statistiek, zie voetnoot 4.

gerelateerd aan feiten (*evidence*) die de betreffende uitspraak wel of niet bevestigen. Omdat het hier primair gaat om de graad van waarschijnlijkheid van onze kennis omtrent de wereld, gebruikt hij het woord epistemisch (onze kennis betreffende). Ook hier een paar voorbeelden ter verduidelijking:
- Wat er kort geleden boven water is gekomen, maakt het onwaarschijnlijk dat de verpleegkundige de dood van haar patiënten op haar geweten heeft. [19]
- Nieuwe astronomische metingen maken het waarschijnlijk dat ons heelal een begin heeft gehad.
Carnap noemt dit de inductieve interpretatie van *probability.* Deze vorm van waarschijnlijkheid is niet los te denken van een persoon die zich, weliswaar op basis van feiten, een beeld vormt van iets in de werkelijkheid.
Een ander voorbeeld dat Mellor noemt, en dat voor ons onderwerp van belang is, betreft de toetsing van een hypothese. Wanneer je bijvoorbeeld besluit om op grond van een steekproefuitkomst (*evidence*) de nul-hypothese te verwerpen, anders gezegd, wanneer de steekproefuitkomst statistisch significant blijkt te zijn, dan doe je die uitspraak op basis van een bepaalde mate van waarschijnlijkheid die van epistemische aard is, en die verschilt van een uitspraak die gebaseerd is op kansberekening. Dit onderscheid doet denken aan het door Deming naar voren gebrachte onderscheid tussen *enumerative studies* en *analytic studies.* Ik zal hierop in de volgende paragraaf terugkomen, wanneer de betekenis van het begrip significantie wederom aan de orde wordt gesteld.
*3. Subjective probability*: dit komt overeen met de bovengenoemde subjectieve kans, het betreft een *personal degree of belief*. Voorbeeld:
Ik acht het niet waarschijnlijk dat Nederland wereldkampioen voetbal zal worden.

Nu we het kansbegrip nader onderzocht hebben, kunnen we de betekenis van het significantiebegrip onder de loep nemen.

## HET SIGNIFICANTIEBEGRIP OPNIEUW BEKEKEN

Bij de interpretatie van de uitkomsten van een significantietoets wordt wel eens over het hoofd gezien dat de huidige statistiek gebaseerd is op het frequentistische kansbegrip. Zoals we hierboven al zagen betekent dit dat we geen kansen kunnen toekennen aan gebeurtenissen die eenmalig zijn.

[19] Naar aanleiding van het geruchtmakende proces tegen Lucia de Berk, de verpleegster die kort geleden is vrijgesproken van de beschuldiging van betrokkenheid bij de dood van een aantal van haar patiënten, kun je je afvragen of het onderscheid tussen *physical* en *epistemic probability* hier relevant is. Wat was het geval? In de nacht van 4 september 2001 sterft in het Haagse Juliana Kinderziekenhuis (JKZ) volkomen onverwacht een baby van een half jaar oud. Artsen denken eerst aan een natuurlijke dood, maar beginnen daar al snel aan te twijfelen: in het bijzijn van Lucia de Berk, de verpleegster die ook in de nacht van 4 september dienst had, overlijden opvallend veel kinderen. In de periode van 343 dagen tot aan 4 september waren er acht kinderen op de afdeling waar zij werkte, in acuut levensgevaar gekomen en steeds had zij dienst. Het grote probleem bij deze zaak was het gebrek aan direct bewijs. De officier van justitie wilde weten of dit alles ook een bizarre speling van het lot zou kunnen zijn. Daartoe werd deze vraag voorgelegd aan een statisticus die op basis van de kansrekening uitkwam op een kans van 1 op de 342 miljoen. Ik ga hier niet op alle details in; zo was de berekening mede gebaseerd op de aanwezigheid van Lucia de Berk op twee andere afdelingen in een ander ziekenhuis, heeft de kansberekening zelf een golf van kritiek van collega-statistici tot gevolg gehad en ook was het niet zo dat de berekende kans door de rechters als wettig en overtuigend bewijs werd opgevat (maar ik acht het niet waarschijnlijk – in subjectieve zin – dat deze kansberekening geen enkele rol heeft gespeeld bij de formulering van het eerste vonnis) . De vraag hoe statistische theorie en werkelijkheid zich tot elkaar kunnen verhouden komt in deze zaak wel erg sterk naar voren en je kunt je afvragen of het hanteren van het onderscheid van Mellor veel ellende had kunnen voorkomen. Lucia de Berk heeft namelijk ongeveer zes jaar vastgezeten.

Laten we ter illustratie hiervan het volgende voorbeeld bekijken.[20]

In het jaar 2000 bleek in een steekproef van 734 gezinnen het gemiddelde aantal kinderen 1.79 te zijn. Men wil nagaan of het gemiddelde aantal kinderen van Nederlandse gezinnen lager ligt dan 2. Mocht deze alternatieve hypothese ( $H_1$ of $H_a$ ) bevestigd worden, dan wijst dat in de richting van een afname van de bevolkingsomvang. De nul-hypothese luidt hier dat het gemiddelde aantal kinderen 2 is. De p-waarde blijkt zeer klein te zijn, namelijk: 0,000003. Op basis hiervan wordt de nul-hypothese verworpen (zelfs bij een significantieniveau van 0,001); het is zeer waarschijnlijk dat het populatiegemiddelde lager ligt dan 2.[21] Tot zover het resultaat van de betreffende berekeningen. Maar wat kunnen we nu wel en wat niet beweren omtrent het gemiddelde aantal kinderen van Nederlandse gezinnen?

Omdat de huidige klassieke statistiek gebaseerd is op de frequentistische benadering van het kansbegrip en, zoals we in de vorige paragraaf gezien hebben, binnen die benadering geen kansen toegekend kunnen worden aan een eenmalige gebeurtenis, luidt de correctie interpretatie als volgt:
wanneer we de procedure van het trekken van een steekproef uit de betreffende populatie en van de daaropvolgende hypothesetoetsing vele malen zouden herhalen (wat we in werkelijkheid natuurlijk niet doen, we trekken slechts één steekproef), dan zouden we de nulhypothese in $\alpha\%$ (het gekozen significantieniveau) van de gevallen ten onrechte verwerpen (O'Hagan & Luce 2003; Howson & Urbach 2006).
Voor een goed begrip van deze interpretatie is het van belang hier op te merken dat we, als we een hypothese toetsen, we telkens conditionele uitspraken doen: àls de nul-hypothese waar is, dàn is de kans enz. Omdat wij niet weten of $H_0$ of $H_1$ waar is, omdat wij niet weten wat de werkelijke toestand van de wereld is, is iedere uitspraak conditioneel op het waar zijn van $H_0$ of $H_1$ (Bijleveld & Commandeur 2009: 36).
Uitspraken als 'de kans dat de de nul-hypothese waar is, is ..%', of 'de kans dat ik een juiste uitspraak doe is ..%', e.d. zijn onzin.[22] In dit verband is het goed om er op te wijzen dat het statistische begrip 'significantie' volgens de grote Van Dale niet 'veelbetekenend' maar 'verantwoorde conclusies toelatend' betekent. Volgens de

[20] Het voorbeeld is ontleend aan Manfred te Grotenhuis & Theo van der Weegen (2008). *Statistiek als hulpmiddel. Een overzicht van gangbare toepassingen in de sociale wetenschappen.* Assen: Van Gorcum, p.63-64.
[21] Het uitvoeren van een significantietoets behoort tot het terrein van de zogeheten inferentiële statistiek,ook wel verklarende, inductieve of mathematische statistiek genoemd. Hierbij wordt op basis van de uitkomsten van een steekproef uitspraken gedaan over de populatie waaruit de steekproef getrokken is. Naast de hypothesetoetsing hoort ook het berekenen van betrouwbaarheidsintervallen (*confidence intervals)* tot het gebied van de inferentiële statistiek. Hierbij wordt op basis van een gevonden steekproefgemiddelde of steekproefproportie een interval berekend dat op de betreffende populatie van toepassing is. Een bekende toepassing hiervan wordt gevormd door de peilingen bij aanstaande verkiezingen.
[22] Iets dergelijks geldt voor een betrouwbaarheidsinterval. Ter illustratie het volgende voorbeeld: in 2009 werd aan ruim 400 personen van 16 jaar en ouder in de provincies Groningen en Drenthe een aantal vragen gesteld waaronder de vraag 'Kan de (economische) crisis leiden tot ontslag van uzelf of van één van uw naasten?'. In die steekproef antwoordt 48 procent van de ondervraagde respondenten bevestigend. Op basis hiervan kan een 95% betrouwbaarheidsinterval van 0.43 – 0.53 berekend worden. De correcte interpretatie luidt: wanneer we de procedure van het nemen van een steekproef (van dezelfde omvang) en het op basis van de steekproefuitkomst berekenen van een 95% betrouwbaarheidsinterval vele malen zouden herhalen, dan zou in 95% van de gevallen de werkelijke populatiefractie (het percentage van alle inwoners van Groningen en Drenthe van 16 jaar en ouder) binnen het berekende betrouwbaarheidsinterval liggen. Ook hier zijn uitspraken als 'de kans dat de populatiefractie tussen 43% en 53% ligt, is 95%', e.d. pertinente onzin.

auteurs van een veelgebruikt leerboek statistiek[23] betekent *significant* niet *important* , maar is het afgeleid van *signifying something* dat vertaald kan worden door 'te kennen geven' of 'op iets duiden'.
Onderzoeksresultaten zijn significant in statistische zin wanneer ze relatief zelden voorkomen bij aselecte steekproeftrekkingen uit een en dezelfde populatie, verondersteld dat de nul-hypothese waar is. In zo'n geval gaan we ervan uit dat toevalsvariaties niet de verklaring vormen voor wat we gevonden hebben, bijvoorbeeld een gevonden verschil tussen een experimentele groep en een controlegroep.
Anders geformuleerd: het toetsen op statistische significantie levert een p-waarde op die de kans is op het verkrijgen van de gevonden steekproefuitkomst of een nog extremere uitkomst, onder de veronderstelling dat de nul-hypothese waar is. Deze p-waarde kan vervolgens gebruikt worden voor de beslissing (vergelijk Fisher's *decision procedure*) om het idee dat de verkregen steekproefuitkomst door toeval tot stand is gekomen, al dan niet te accepteren.

Het is de vraag in welke mate het kansbegrip in de onderzoekspraktijk correct geïnterpreteerd wordt. Wat medische onderzoekers betreft (in de context van *evidence-based medicine)* schrijft Goodman: *"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with* $P = 0.05$*, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect. This is an understandable but categorically wrong interpretation because the* $P$ *-value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true."* (Goodman 1999: 997-998). De laatste zin uit dit citaat is interessant: een verkeerde interpretatie van het kansbegrip is debet aan het toekennen van een beslissende rol van de steekproefgegevens.

Dit alles neemt niet weg dat we naar mijn mening op basis van een uitgevoerde significantietoets wel degelijk goede gronden kunnen hebben om aan te nemen dat iets wel of niet het geval zal zijn in de werkelijkheid. Alleen kunnen we de onzekerheid die hierbij komt kijken blijkbaar niet op dezelfde wijze uitdrukken als in het geval van *enumerative statistical studies* (Deming) en *physical or statistical probability* (Mellor). Dit brengt ons bij de vraag of statistische toetsingsresultaten op zich wel een voldoende basis vormen voor het doen van wetenschappelijk verantwoorde uitspraken over de werkelijkheid. Dit is het onderwerp van de volgende en laatste paragraaf.

## PERIKELEN ROND DE P-WAARDE

Er zijn twee manieren om theorieën en hypothesen enerzijds en de waargenomen feiten anderzijds op elkaar te betrekken, namelijk langs deductieve en langs inductieve weg.[24] In het eerste geval beginnen we met het formuleren van een hypothese en leiden we daaruit af wat in de werkelijkheid het geval zal zijn. Het voordeel van deze weg is dat ze

[23] D.S.Moore, G.P.McCabe & B.A.Craig (2009). *Introduction to the Practice of Statistics* (6th Rev. ed.) New York: Freeman, p.379.
[24] De volgende beschouwingen zijn voor een belangrijk deel ontleend aan Goodman (1999).

objectief is, dat wil zeggen dat onze uitspraken over de werkelijkheid waar zullen zijn wanneer de onderliggende hypothese of theorie waar is. Het nadeel is dat onze kennis beperkt blijft tot hetgeen in de hypothese is geformuleerd.
Bij de inductieve weg gaan we omgekeerd te werk, op basis van waargenomen feiten bepalen we welke hypothese het meest aannemelijk is. Het concept van beschikbare *evidence* is inductief van aard. Het grote voordeel van deze weg is dat we op het spoor kunnen komen van nieuwe hypothesen en theorieën en op die manier nieuwe inzichten kunnen verwerven. Het nadeel is dat we er niet zeker van kunnen zijn dat wat we omtrent de werkelijkheid concluderen ook feitelijk zo is. Dit gaat terug op het inductieprobleem dat Karl Popper illustreerde aan de hand van het voorbeeld van de witte zwanen.

Als we terug in de tijd gaan kunnen we de volgende zaken op een rijtje zetten. We begonnen onze beschouwing met het bespreken van de logisch positivisten die wetenschap en de ontwikkeling van wetenschap primair vanuit de inductieve weg beschreven. Popper bekritiseerde deze visie omdat we volgens hem we altijd vanuit een theoretische context naar de werkelijkheid kijken, er bestaan geen naakte feiten. Wel moeten we volgens Popper altijd trachten onze theorieën en hypothesen te falsifiëren, we moeten zogezegd altijd kritisch blijven ten aanzien van de ratio, vandaar dat zijn visie als kritisch rationalisme bestempeld wordt.
Teneinde ook statistische hypothesen falsifieerbaar te maken, ontwikkelde Ronald Fisher de theorie van de p-waarde die gezien kan worden als een methode om de kracht van de data (*evidence*) te bepalen.
Al snel kwam er kritiek op de praktische bruikbaarheid van het concept van de p-waarde. Het blijkt namelijk zo te zijn dat de p-waarde beïnvloed wordt door de steekproefomvang. Het praktische gevolg is dat een heel klein effect bij een grote steekproef dezelfde p-waarde kan opleveren als een groot effect in een kleine steekproef.[25] Deze kritiek heeft er toe geleid dat het accent verschoof van het concept van de p-waarde naar de berekening van betrouwbaarheidsintervallen (zie voetnoot 21).
Tegelijkertijd zochten de statistici Neyman en Pearson naar een manier om wiskundige formules toe te kunnen passen binnen de statistiek. Zij vonden die in hun concept van de hypothesetoetsing waarbij de berekende kansen conditioneel zijn op de waarheid van de nul-hypothese. Dit is in feite een deductieve en daardoor objectieve (in de zin zoals hierboven aangegeven) benadering.[26] Neyman en Pearson realiseerden zich dat het concept van de hypothese toetsing een procedure is die het mogelijk maakt om een hypothese te accepteren of the verwerpen maar geen methode om de *strength* van de data te bepalen. Daardoor was en is de hypothesetoetsing geen inductieve methode om op basis van een steekproef de waarschijnlijkheid van een nul-hypothese te berekenen,

[25] Bijleveld en Commandeur (2009: 36-37) geven het volgende voorbeeld: "Als een steekproef echter fors groot wordt, is bijna ieder minuscuul verband of verschil significant. Significantietoetsing heeft dan geen zin meer: alles is significant en toetsing is daarmee nietszeggend geworden. Bij een steekproef van 200.000 respondenten slaat bijvoorbeeld een correlatiecoëfficiënt ter waarde van 0.01 nog uit als significant verschillend van 0. Het zij duidelijk dat dat verband in absolute zin te klein is om van een betekenisvolle samenhang te spreken."
[26] Het element van de persoonlijke of, zo u wilt, subjectieve inschatting van het belang van de data voor de waarschijnlijkheid van de betreffende hypothese, dat bij de inductieve weg altijd een rol speelt vanwege het element van de onzekerheid, zagen Neyman en Pearson in de keuze van de onderzoeker ten aanzien van de fouten van de eerste en de tweede soort. Meer concreet gaat het om de vraag welke fout als relatief ernstig gezien wordt. Bij sociaal-wetenschappelijk onderzoek wordt de eerste fout doorgaans ernstiger gevonden dan de tweede (Bijleveld & Commandeur 2009).

maar een deductieve methode om het aantal onjuiste conclusies (*errors*) bij een groot aantal herhalingen van het kansexperiment te minimaliseren. Vandaar dat de huidige statistiek ook wel als *error-based* gekwalificeerd wordt. In hun eigen woorden:
*"no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.*
*(…)*
*But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we ensure that, in the long run of experience, we shall not often be wrong."* [27]

Zoals we eerder zagen, heeft de p-waarde van Fisher betrekking op de vraag: hoe waarschijnlijk is het dat de nulhypothese waar is, gezien de data (steekproefuitkomst)?, terwijl de Neyman-Pearson methode van hypothesetoetsing een antwoord geeft op de vraag: is de p-waarde kleiner dan of gelijk aan $\alpha$ ? Op grond van de p-waarde kan, gezien het frequentistische karakter van het gehanteerde kansbegrip, geen kansuitspraak over een enkele nulhypothese gedaan worden. Bovendien wees Fisher op het probleem van de 'oneindige hypothetische populatie'. Het gevolg is dat we op basis van een p-waarde alleen uitspraken kunnen doen met het karakter van een *epistemic probability* (Mellor 2005) zoals bijvoorbeeld: 'Gezien de data is het onwaarschijnlijk dat de nulhypothese waar is'. Deze uitspraak verschilt wezenlijk van een uitspraak als 'de kans dat ik twee keer een zes gooi wanneer een zuivere dobbelsteen 10 keer geworpen wordt, is ongeveer 0.2907' (*objective propability).*
Het toepassen van de Neyman-Pearson methode van hypothesetoetsing daarentegen, resulteert in een beslissing: de nulhypothese wel of niet verwerpen en resulteert niet in een kansuitspraak. Het enige kansaspect hierbij is dat je kunt zeggen dat je op lange termijn in $\alpha$ % van alle gevallen een verkeerde beslissing neemt.

Het is vergelijkbaar met de volgende situatie. Een rechter heeft in een groot aantal strafzaken, bijvoorbeeld 1000, een vonnis uitgesproken. De rechter kan met betrekking tot een individuele verdachte, zeg verdachte A, niet met zekerheid zeggen of A schuldig of onschuldig is. De rechter kan ook geen kansuitspraak over een individuele verdachte doen zoals 'de kans dat verdachte A schuldig is, is 95% '. Het enige (objectieve) kansaspect zit hem hierin dat je kunt zeggen dat de rechter in $\alpha$ % van de 1000 strafzaken een onjuist vonnis zal uitspreken. De rechter kan ten aanzien van een individuele verdachte wel zeggen: gezien de feiten (lees: de gevonden p-waarde) acht ik het zeer waarschijnlijk dat verdachte A schuldig is. Dit betreft dan een *epistemic probability.* Zo'n uitspraak is dan gebaseerd op de betreffende feiten in samenspraak met eerder verworven theoretische inzichten, praktijkervaring, professionele intuïtie enz. De vergelijking gaat echter mank omdat in geval van onderzoek het slechts een enkele steekproef of experiment betreft. De 999 herhalingen van de procedure van steekproeftrekking zijn dus louter hypothetisch en niet feitelijk.

Hier is naar mijn mening een spanningsveld ontstaan. In plaats van een methode om op basis van een enkel experiment of ander steekproefonderzoek uitspraken te kunnen doen over een hypothese hebben Neyman en Pearson, vanuit hun streven naar objectiviteit,

[27] Het citaat is ontleend aan Goodman (1999): 998.

de methode van significantietoetsing ontwikkeld die slechts het aantal onjuiste conclusies op de lange termijn wil minimaliseren. Maar dat druist in tegen de wetenschappelijk intuïtie die zegt dat we moeten proberen om op basis van ons (eenmalig) onderzoek iets zinnigs te zeggen over de werkelijkheid. En dat is zeer wel mogelijk, maar alleen in epistemische zin.
Het gaat al met al om de vraag wat voor soort uitspraak je als onderzoeker wilt doen op basis van de verkregen data. Wanneer je gebruik wilt maken van het objectieve kansbegrip moet je genoegen nemen met een *repetitive error rate*: bij een veelvuldige herhaling van mijn onderzoek zal ik in $\alpha$ % van de gevallen een verkeerde beslissing nemen. Wanneer je echter iets wilt zeggen over de betekenis of (bewijs)kracht van de verkregen data, hetgeen vermoedelijk doorgaans het geval zal zijn, moet je genoegen nemen met een epistemische waarschijnlijkheid. Dit alles betekent dat we bij de formulering van uitspraken over de werkelijkheid op basis van kwantitatief empirisch onderzoek theoretische en/of epistemologische overwegingen enerzijds en resultaten van significantietoetsing anderzijds zorgvuldig op elkaar dienen te betrekken. Een voorbeeld van een overweging van epistemologische aard betreft het feit dat de ideeën die ten grondslag liggen aan de homeopathie niet verenigbaar zijn met de bestaande natuurkundige wetten.[28]

## CONCLUSIES

Het moge duidelijk zijn geworden dat achter het concept van de significantietoets, zogezegd achter elke knop in SPSS of welke statistische software dan ook, een wereld van beschouwingen van velerlei aard schuil gaat en dat de betekenis van dit concept complexer is dan over het algemeen wordt aangenomen. Dit artikel is dan ook bedoeld als een pleidooi voor een reflexief gebruik van statistische methoden en software en tegen een al te instrumentele opvatting van statistische toetsing en, in dat verband, als een waarschuwing voor "een perverse prikkel die onderzoekers aanzet om zo veel mogelijk significante resultaten bij elkaar te vissen" (Katan 2010). We moeten erop bedacht zijn dat significantietoetsing geen basis levert voor een inductieve gevolgtrekking (*inference)* maar slechts een gedragsrichtlijn oplevert: het verwerpen dan wel accepteren van een hypothese. De p-waarde wekt echter de schijn een *measure of evidence* op basis van een enkele steekproef in te houden, reden om haar vaak aan het begin van een onderzoeksrapport te vermelden terwijl de conclusies pas later aan bod komen. Kortom, de p-waarde krijgt over het algemeen een te grote (en onjuiste) rol toebedeeld. Carver (1993) stelt daarom voor om in onderzoeksrapporten eerst de resultaten van de significantietoets te interpreteren en pas daarna de betreffende p-waarde te vermelden.

Op grond van het bovenstaande kom ik tot de conclusie dat wanneer we wetenschappelijk verantwoorde uitspraken over de werkelijkheid willen doen, we ons gesteld zien voor de taak om statistische resultaten in samenhang met theoretische overwegingen, met kennis gebaseerd op ervaring en eerder verricht onderzoek en met professionele intuïtie te zien. Anders gezegd, in de context van een proces van

[28] Zie bijvoorbeeld S.Singh & E.Ernst (2008). *Trick or Treatment. The Undeniable Facts about Alternative Medicine.* New York: Norton.

oordeelsvorming.

In de context van het hoger beroepsonderwijs waarin ik zelf werkzaam ben en waar onderzoek steeds centraler komt te staan , zou ik willen pleiten voor het voeren van een discussie over de vraag welke statistische methoden het beste aansluiten bij de specifieke onderzoeksvragen en het type onderzoek binnen een opleiding of beroepenveld, alsmede over de vraag hoe de juiste balans gevonden kan worden tussen statistische methoden en theorievorming. Het is niet ondenkbaar dat de nadruk die al geruime tijd op significantietoetsen gelegd wordt, ten koste is gegaan van theoretische overwegingen en van theorievorming in het algemeen. In dat verband verdient het aanbeveling om in de toekomst binnen het (beroeps)onderwijs meer aandacht te besteden aan het uitwerken van de relatie tussen theoretische concepten en empirische analyses.

augustus 2010
Stenden hogeschool Leeuwarden

## LITERATUUR

Bijleveld, C.C.J.H., & Commandeur, J.J.F. (2009). *Multivariate analyse. Een inleiding voor criminologen en andere sociale wetenschappers.* Den Haag: Boom.

Carver, R.P. (1978). The Case Against Statistical Significance Testing. *Harvard Educational Review*, 48(3), 378-399.

Carver, R.P. (1993). The Case Against Statistical Significance Testing, Revisited. *The Journal of Experimental Education,* 61(4), 287-292.

Cohen, L.J. (1989). *An Introduction to the Philosophy of Induction and Probability.* Oxford: Clarendon Press.

Cowles, M., & Davis, C. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist*, 37(5), 553-558.

Deming, W.E. (1975). On Probability As a Basis For Action. *The American Statistician,29(4), 146-152.*

Dooremalen, H., De Regt, H., & Schouten, M. (2007). *Exploring Humans. An Introduction to the Philosophy of the Social Sciences.* Amsterdam: Boom.

Groot, A.D. de (1975). *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. '*s-Gravenhage: Mouton.

Hacking, I. (2006). *The Emergence of Probability. A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference* (2nd ed.). New York: Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning. The Bayesian Approach* (3rd ed.). Chicago: Open Court.

Gillies, D. (2000). *Philosophical Theories of Probability.* London/New York: Routledge.

Goodman, S.N. (1999). Toward Evidence-Based medical Statistics. 1: The *P* Value Fallacy. *Annals of Internal Medicine, 130(12),* 995-1004.

Katan, M. (2010, 15 mei). Overheersende toevalstreffers. *NRC,* p. Wetenschap 2.

Kerridge, D., & Kerridge, S. (1998). *Statistics and Reality.* Retrieved April 3, 2010, from http://en.wikipedia.org, *Analytic and enumerative statistical studies,* external link, pdf.

Klep, P.M.M. (2002). A Historical Perspective on Statistics and Measurement in the Netherlands 1750-1850. In P.M.M. Klep & I.H. Stamhuis (Eds), *The Statistical Mind in a Pre-Statistical Era: The Netherlands 1750-1850* (p.29 – 69). Amsterdam: Aksant.

Meester, R. (2004). *Het pseudoniem van God. Een wiskundige over geloof, wetenschap en toeval.* Ten Have.

Lucas, J.R. (1970). *The Concept of Probability.* Oxford: Clarendon Press.

Mellor, D.H. (2005). *Probability: A Philosophical Introduction.* London/New York: Routledge.

O'Hagan, A., & Luce, B.R. (2003). *A Primer on Bayesian Statistics in Health Economics and Outcomes Research.* Retrieved June 25th , 2010, from http://www.shef.ac.uk/chebs/.

Salsburg, D. (2002). *The Lady Tasting Tea. How Statistics Revolutionized Science in the Twentieth Century.* New York: Holt.

Popper, K.R. (1975). *The Logic of Scientific Discovery.* (8th rev. ed.). London: Hutchinson.

Stigler, S.M. (1996). The History of Statistics in 1933. *Statistical Science, Vol. 11, No. 3,* 244-252.

Stigler, S.M. (2003). *The History of Statistics. The Measurement of Uncertainty before 1900.* Cambridge, Massachusetts: The Belknap Press of Havard University Press.

Tijms, H. (2007). *Understanding Probability. Chance Rules in Everyday Life.* New York: Cambridge University Press.

Wagenmakers, E., Lee, M., Lodewyckx, T., & Iverson, G.J. (2008). Bayesian Versus Frequentist Inference. In H. Hoijtink, I. Klugkist, & P.A. Boelen (Eds), *Bayesian Evaluation of Informative Hypotheses* (p. 181-207). New York: Springer.