Small step or giant leap

Op de weg naar transparante (subjectieve) wetenschap

Mijnheer de Rector Magnificus,

Zeer gewaardeerde collega's, lieve familie-leden, vrienden en overige toehoorders.

Ik mag 45 minuten praten! Om jullie enig houvast te geven bij het luisteren, ziet u hierbij de indeling die mij houvast heeft gegeven bij het schrijven.

Outline

1 Kindertijd
2 P's and Bayes
3 Dogs and horses
4 Data Science
5 Excuses
6 Vrouwen
7 Dank & drank

Zoals u ziet beginnen we bij het begin, mijn kindertijd. Het is me namelijk opgevallen dat bij veel van dit soort toespraken wordt gestart met een terugblik op de kinderjaren. In die kinderjaren zijn er dan blijkbaar vaak al allerlei tekenen van een wetenschappelijke nieuwsgierigheid, die als voorbode gezien kunnen worden van de wetenschappelijke loopbaan van de spreker.

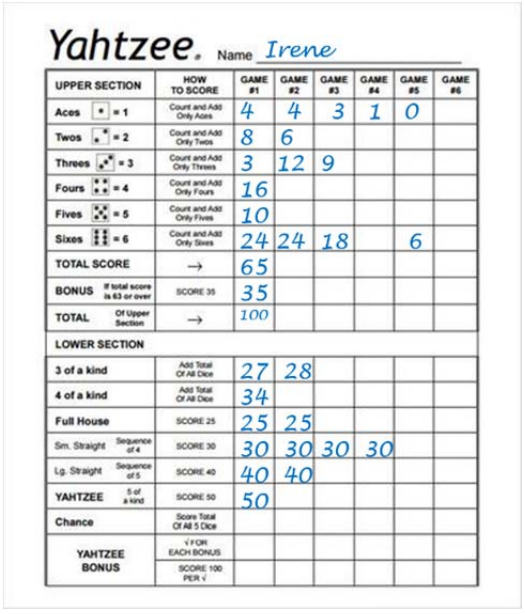
Nu houd ik me bezig met het verbeteren en ontwikkelen van data analyse technieken. Hierbij horen zaken als het schatten van effecten of relaties, met als kern het inschatten van onzekerheden bij die schattingen, evenals het inschatten van de kansen op fouten als hieruit bepaalde conclusies getrokken worden. Was ik als kind dan al geboeid door kansen of kansberekening?

Ik weet vrij zeker dat het antwoord daarop “nee” is en dat ik mijn kindertijd niet gehinderd door al te veel onzekerheden ben doorgekomen en me dan ook niet bezighield met kansen op bepaalde gebeurtenissen of kansen om onjuiste beslissingen te maken.

Wel denk ik dat mijn ouders zullen beamen dat ik al van jongs af aan heel kritisch kan zijn. Bijvoorbeeld kritisch op het correct formuleren van zaken. Hoe vaak mijn vader ook uitspraken deed als: “wat zijn die spruitjes weer lekker”, of “wat is dit toch een prachtige opera” ik was nooit te beroerd om hem te verbeteren met de standaard reactie “nee, jij vindt spruitjes lekker” en “jij vindt die opera mooi”. Het moge duidelijk zijn dat ik z'n smaak in groenten en muziek niet deelde. En dat is trouwens nog steeds zo.

Kritisch zijn, onder anderen in het correct formuleren en interpreteren van wat onderzoeksresultaten *wel* maar vooral ook *niet* betekenen, is een belangrijk aspect van wetenschappelijk werk. Dit bespreken en stimuleren in zowel onderwijs als onderzoek zie ik als een van mijn taken als methodoloog en statisticus. Ik zal hier later nog op terug komen.

Nu eerst nog even terug naar mijn kindertijd en kansen en toeval. Volgens mij hield ik er juist niet zo van. Zo herinner ik me dat we als gezin een paar zomers op rij nogal slecht weer hadden op vakantie en urenlang in de caravan yahtzee hebben gespeeld. Nu is yahtzee een dobbelspel en daarmee heeft pech en geluk, ofwel toeval, een grote rol. Daarnaast is enig tactisch inzicht ook van belang, omdat je veel worpen in meerdere vakjes kwijt kan en dus verstandige keuzes moet maken.



Yahtzee. Name Irene

UPPER SECTION	HOW TO SCORE	GAME #1	GAME #2	GAME #3	GAME #4	GAME #5	GAME #6
Aces = 1	Count and Add Only Aces	4	4	3	1	0	
Twos = 2	Count and Add Only Twos	8	6				
Threes = 3	Count and Add Only Threes	3	12	9			
Fours = 4	Count and Add Only Fours	16					
Fives = 5	Count and Add Only Fives	10					
Sixes = 6	Count and Add Only Sixes	24	24	18		6	
TOTAL SCORE	→	65					
BONUS If total score is 63 or over	SCORE 35	35					
TOTAL Of Upper Section	→	100					
LOWER SECTION							
3 of a kind	Add Total Of All Dice	27	28				
4 of a kind	Add Total Of All Dice	34					
Full House	SCORE 25	25	25				
Sm. Straight Sequence of 4	SCORE 30	30	30	30	30		
Lg. Straight Sequence of 5	SCORE 40	40	40				
YAHTZEE 5 of a kind	SCORE 50	50					
Chance	Score Total Of All 5 Dice						
YAHTZEE BONUS	√ FOR EACH BONUS						
	SCORE 100 PER √						

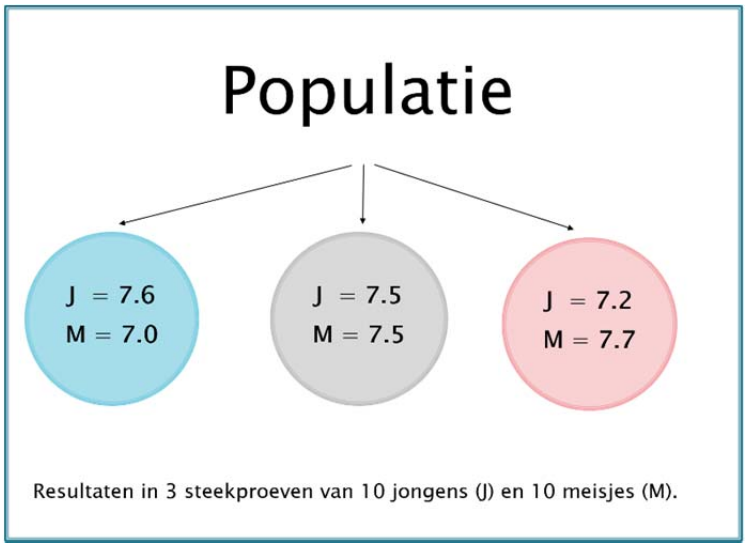
Nu weet ik niet wie in ons gezin het toen bedacht heeft maar wij speelden meestal niet per kolom, dus een enkel spel, maar per kaart. Ofwel, een spel was pas klaar en gewonnen of verloren als alle 6 kolommen gevuld waren. Wat we daarmee feitelijk deden was de rol van het toeval in het bepalen van de winnaar kleiner maken en de rol van slim, tactisch spelen groter. Ik weet zeker dat ik toen met geen mogelijkheid had kunnen uitleggen waarom dit zo is. Ik was een jaar of 10.

Maar nu, nu ik iets meer begrijp van statistiek en kansen, is het eigenlijk wel een mooie brug naar het praten over kleine en grote aantallen en de invloed die dat heeft op toevallige resultaten.

We doen onderzoek om nieuwe kennis te verkrijgen. Ik zou me bijvoorbeeld kunnen afvragen of meisjes en jongens een even goed ruimtelijk inzicht hebben. Om dit te onderzoeken moet je eerst heel duidelijk maken wat je precies bedoelt met ruimtelijk inzicht en hoe je dit gaat meten. Ik ben er voor het gemak van uitgegaan dat ruimtelijk inzicht is gemeten met rapportcijfers, dus van 0-10, waarbij een 10 de beste score is.

Daarnaast moet je bepalen over wie je het precies wilt hebben, zoals bijvoorbeeld een afbakening tot basisschool leerlingen in Nederland. Vervolgens zal je zelden alle kinderen die binnen die afbakening vallen onderzoeken; we nemen meestal een steekproef uit de beoogde populatie. Laten we er voor het gemak even vanuit gaan dat dit een helemaal goed getrokken steekproef is, dus zonder verstorende selectie effecten. En dat we heel goed het beoogde aspect, in dit geval ruimtelijk inzicht, hebben gemeten. Hoe goed kun je dan je vraag over een populatie beantwoorden met slechts deze steekproef?

In deze figuur is te zien dat dit voor kleine steekproeven helemaal niet zo goed gaat.



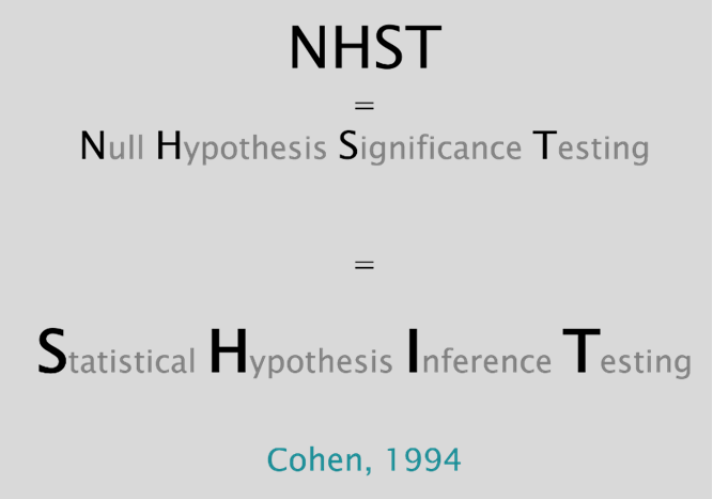
Resultaten in 3 steekproeven van 10 jongens (J) en 10 meisjes (M).

Als ik drie keer een steekproef uit dezelfde populatie trek en geinteresseerd ben in het verschil in ruimtelijk inzicht tussen jongens en meisjes, krijg ik niet alleen drie keer een ander getal maar ook drie keer een andere conclusie, namelijk jongens scoren hoger, het resultaat in de blauwe cirkel, meisjes scoren hoger, het resultaat in de roze cirkel, of er is geen of nauwelijks verschil tussen jongens en meisjes, de grijze cirkel. De variatie in uitkomsten tussen verschillende steekproeven uit eenzelfde populatie kan dus behoorlijk groot zijn. De eenvoudigste oplossing om deze verschillen kleiner te maken is het nemen van grotere steekproeven. Door meer mensen te onderzoeken krijg je meer stabiele resultaten en dus ook meer zekerheid dat je steekproefresultaat dicht bij het werkelijke populatie effect zit. Want wat we willen weten was tenslotte niet of de gemiddelden van jongens en meisjes in de bestudeerde steekproef verschillend zijn maar vooral of het steekproefresultaat een indicatie is dat er in de beoogde populatie een verschil is. Om hier een antwoord op te kunnen geven rapporteren we vrijwel standaard de p-waarde, die gebruikt wordt om te bepalen of een nulhypothese die stelt dat er geen verschil is wel of niet wordt verworpen.

De p-waarde wordt ook wel de significantie genoemd, hoewel dit mensen ernstig op het verkeerde been kan zetten over wat het betekent. Een synoniem voor het woord significant is betekenisvol of belangrijk. Hiermee zou je kunnen concluderen dat een significant verschil in ruimtelijk inzicht tussen jongens en meisjes daarmee ook een belangrijk of betekenisvol verschil is. Dit is echter geen juiste conclusie. Tevens wordt een niet-significant resultaat wel geïnterpreteerd als bewijs voor het afwezig zijn van een verschil. Ook dit is een verkeerde conclusie. Zoals in mijn introductie al gezegd: het is uitermate belangrijk om zaken correct te formuleren om het trekken van foutieve conclusies te vermijden.

Voor de hier aanwezige wetenschappers is dit zeer basale informatie. Dit is namelijk wat we onze studenten al in hun eerste studiejaar leren. Wat we onze studenten niet standaard leren, hoewel dat in een wetenschappelijke opleiding misschien best zou mogen, is dat er een heleboel statistici zijn, die de p-waarde maar helemaal niks vinden. Dit blijkt bijvoorbeeld uit het volgende:

De meest bekende afkorting voor het toetsen van hypothesen met behulp van p-waarden is NHST. Dit staat voor Null Hypothesis Significance Testing. Deze term wordt onder anderen door de statisticus Jacob Cohen gebruikt, hoewel hij ook heeft aangegeven ernstig te hebben getwijfeld over de benaming Statistical Hypothesis Inference Testing...



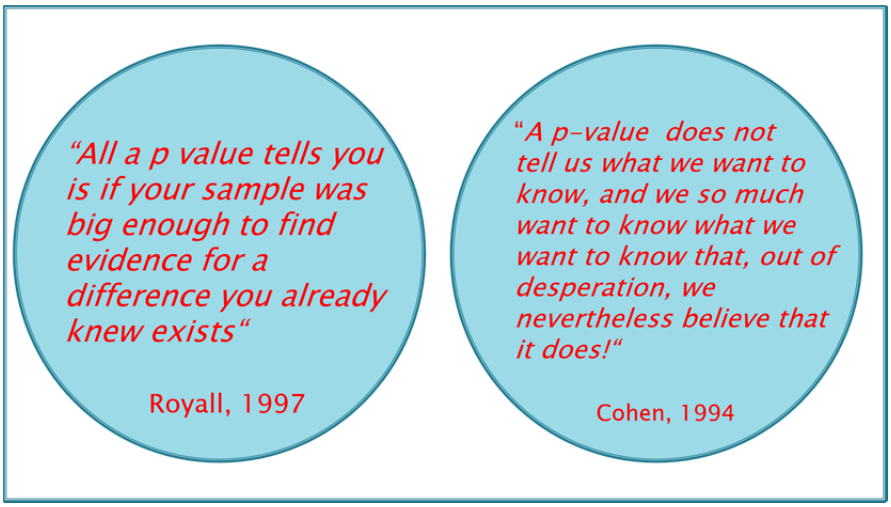
Een veel besproken kritiek op nulhypothese toetsen is dat de nulhypothese, die stelt dat het verschil tussen 2 gemiddelden in een populatie, zoals in ons voorbeeld van ruimtelijk inzicht, exact 0.00000 is, heel onwaarschijnlijk is, of zelfs onmogelijk. Richard Royall stelt vanuit dat argument, het volgende:

"All a p value tells you is if your sample was big enough to find evidence for a difference you already knew exists".

Ofwel, je weet strikt genomen al, dat het onmogelijk is dat het verschil in twee gemiddelden in de populatie exact nul is. Er is dus altijd wel een verschil, mogelijk een extreem klein –en wellicht onbelangrijk- verschil. Royall's belangrijkste punt is dan ook dat hij liever een methode voor het toetsen van hypothesen gebruikt, die interessante en realistische hypothesen kan evalueren. De nulhypothese doet hij af als niet plausibel en dus niet interessant.

Wat je ook in z'n quote terugziet is het belang van de grootte van de steekproef. Een p-waarde hangt sterk af van de hoeveelheid data waarop deze is berekend. Daarom kan een p-waarde alleen, nooit iets zeggen over de belangrijkheid van een resultaat. Een piepklein, irrelevant verschil kan, als de steekproef groot genoeg is, zomaar een hele kleine p-waarde opleveren en dus statistische significantie.

De correcte interpretatie van een p-waarde is niet heel ingewikkeld maar misschien ook niet heel intuïtief of informatief. Als onderzoekers studies doen om bepaalde hypothesen te toetsen, dan zou een informatief en makkelijk te interpreteren resultaat iets zeggen over de kans dat elk van je hypothesen waar is. Dat is niet wat de p-waarde je vertelt. Ook hier gebruik ik graag een quote.



Jacob Cohen stelde hierover het volgende: "*a p-value does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!*"

Hiermee wil Cohen benadrukken dat zelfs al weten we wat de correcte interpretatie van een p-waarde is, en dat is dus *niet* de kans dat een bepaalde hypothese waar is, dan nog is het erg verleidelijk om de p-waarde zodanig te interpreteren, omdat dat nu eenmaal is wat we graag willen weten.

Er is een alternatief voor de p-waarde waarmee je wel resultaten krijgt die iets zeggen over de kans dat elk van je hypothesen waar is. Dit is namelijk informatie die de Bayesiaanse statistiek je kan geven.

Voor de niet-statistici onder ons: de p-waarde kun je zien als het basis ingrediënt van wat nog steeds de mainstream voor data analyse is en wordt aangeduid met de term klassieke of frequentistische statistiek. Een nieuwere aanpak die in opkomst is maar ook door velen nog met argusogen bekeken wordt is de zogeheten Bayesiaanse statistiek, vernoemd naar Thomas Bayes.

THOMAS BAYES
1702-1761

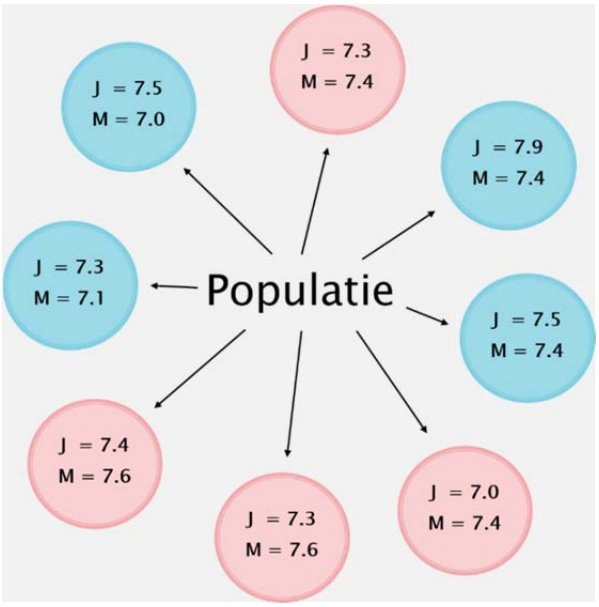
Persoonlijk ben ik erg gecharmeerd van de vele mogelijkheden die de Bayesiaanse aanpak biedt en ik denk dat onderzoekers zichzelf benadelen als ze die aanpak bij voorbaat afwijzen. Maar je zult mij ook niet horen beweren dat Bayes de oplossing voor alles is, of per definitie altijd beter.

Het is naar mijn mening echter wel van onschatbare waarde dat de Bayesiaanse statistiek zijn plek begint te vinden in de gereedschapskist van steeds meer onderzoekers, ook in de sociale en gedragswetenschappen. Dit maakt dat onderzoekers zich meer en meer realiseren dat ze ook voor de analyse van hun data keuzes kunnen en moeten maken. En dat elke methode zijn voors en tegens heeft. Geen enkele methode kan ons objectief, op onbetwistbare wijze, en met zekerheid vertellen hoe het zit. Daarnaast is het zo dat elke onderzoeker binnen elke methode verschillende subjectieve keuzes maakt. Veelal gebeurt dit zonder goed te overzien welke consequenties die keuzes voor de conclusies van het onderzoek hebben. Binnen de klassieke statistiek kun je daarbij denken aan zaken als hoe we omgaan met extreme waarden, missende waarden, geschonden model assumpties, of p-waarde correcties bij meervoudig toetsen. Bayesiaans is het meest voor de hand liggende voorbeeld het formuleren van prior verdelingen en de mogelijke invloed daarvan op de resultaten.

Er wordt gestreefd naar objectiviteit in wetenschap en dat is uiteraard heel belangrijk. Maar echt objectief onderzoek uitvoeren is naar mijn mening een illusie en ik denk dat we dat misschien nog onvoldoende geaccepteerd hebben. En zolang we denken dat we zonder subjectieve keuzes het onderzoeksproces doorlopen, zullen we waarschijnlijk te weinig transparant rapporteren.

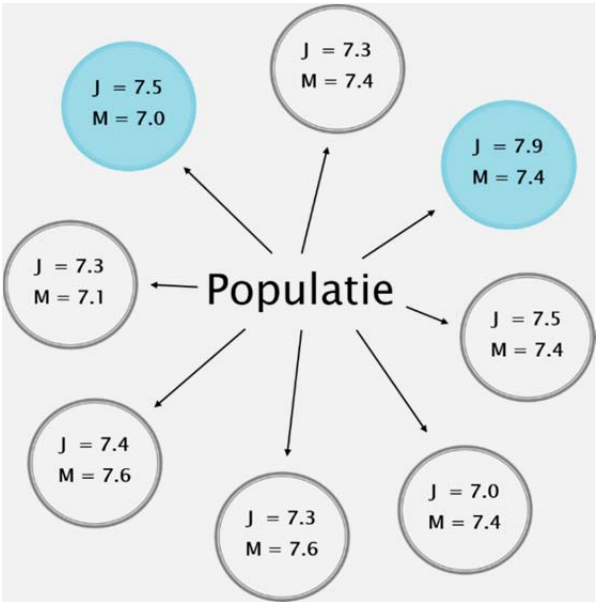
Dit is waarom ik enigszins provocerend vandaag als titel heb gekozen dat we wat mij betreft op de weg zijn naar transparante maar toch ook subjectieve wetenschap. Of dit slechts een klein stapje is, of een enorme sprong ten opzichte van de huidige praktijk, daar ben ik niet helemaal zeker van.

Inmiddels ben ik wat afgedwaald van een eerder gepresenteerd aspect, namelijk de onzekerheid die er is doordat we slechts een steekproef hebben, om te leren en uitspraken te doen over een grotere populatie. Daar wil ik het betoog graag weer oppakken en wel met de volgende figuur.



In deze figuur heb ik 8 herhaalde steekproeven uit precies dezelfde populatie weergegeven, waarbij grijs betekent dat er helemaal geen verschil werd gevonden (iets wat in de steekproeven niet voorkomt), roze dat de meisjes hoger scoorden, en blauw dat de jongens hoger scoorden. De verschillen tussen jongens en meisjes variëren van 0.1 tot 0.5 punten, waarbij in de helft van de studies de jongens en in de andere helft de meisjes hoger scoren.

Als we zoals gebruikelijk op grond van de p-waarde beslissen welke verschillen wel en niet significant zijn, en dat interpreteren als bewijs voor een verschil in de populatiegemiddelden indien p<5% is, of als indicatie dat er geen verschillen zijn in de populatie wanneer p>5%, zien dezelfde resultaten er zo uit.



In 6 studies wordt geen statistisch significant verschil gevonden en in 2 studies scoren jongens gemiddeld significant beter dan meisjes.

Verschillende steekproeven lijken dus wederom tot verschillende conclusies te leiden. Dat is problematisch. Hoe weten we nu wat de waarheid in de populatie is? Hebben jongens een beter ruimtelijk inzicht omdat ik dat twee maal met statistische significantie heb aangetoond? Of is het waarschijnlijker dat er geen verschil is, omdat er in 8 studies 4x uitkwam dat meisjes gemiddeld wat hoger scoren, 4x dat jongens wat hoger scoren en het 6 van de 8 keer bovendien geen statistisch significante verschillen waren.

Laat me als eerste opmerken dat er statistisch gezien niets fout gaat, zolang de conclusies correct geformuleerd worden. Van kleine steekproeven kun je best wat leren maar er is ook nog veel onzekerheid en daarom is het uitermate belangrijk om standaard ook informatie over die onzekerheid te rapporteren.

Onder anderen Geoff Cumming pleit voor het vervangen van p-waarden en geschatte effecten door zogenaamde schattingsintervallen, waarin zowel de schatting van het effect als de onzekerheid bij die schatting is meegenomen.

In deze figuur ziet u per steekproef het geschatte verschil D en een schattigsinterval, weergegeven door de onder- en bovengrens van de schatting. Als we de onzekerheid van de schatting van een effect interpreteren als de breedte van het schattingsinterval, zie je hier dat er een marge van bijna een vol punt is. Als we het grootst gemeten steekproefverschil van een half punt daar mee vergelijken is dat misschien wel heel weinig evidentie voor een verschil tussen jongens en meisjes. Zeker als we ook steekproeven hebben waar het verschil zelfs maar 0.1 punt is.

Rapportage en samenvatting

P	D (J–M)		ondergrens	bovengrens
> .05	−0.1		−0.55	+0.35
< .05	+0.5		+0.05	+0.95
> .05	+0.1		−0.35	+0.55
> .05	−0.4		−0.85	+0.05
> .05	−0.3		−0.75	+0.15
> .05	−0.2		−0.65	+0.25
> .05	+0.2		−0.25	+0.65
< .05	+0.5		+0.05	+0.95

Gemiddelde verschil J–M (ES) en schattingsinterval

Alle studies: ES = +0.04 (−0.12; +0.20)

Significante studies: ES = +0.50 (+0.18; +0.82)

Als we voor deze 8 steekproeven focussen op de significantie en dus of het effect wel of niet is aangetoond, zitten we met tegenstrijdige uitkomsten. Als we kijken naar de onder- en bovengrenzen van de schattingsintervallen zien we dat de resultaten helemaal niet zo heel erg uiteenlopend zijn. Het gaat dus "fout" als we te vergaande conclusies willen trekken uit 1 of meer kleine studies en dat doen zonder duidelijke vermelding van de onzekerheid van de resultaten en de conclusie.

Dit belang wordt nog wel eens onderschat wat blijkt uit ofwel niet rapporteren over de onzekerheid, ofwel het wel rapporteren maar vervolgens nauwelijks bij de interpretatie van de resultaten mee te nemen. Onderzoekers voeren onderzoek uit om kennis te vergaren. Het lijkt erop dat het verleidelijk is om vooral de bevindingen en niet de bijbehorende onzekerheden te benadrukken.

Dit voorbeeld waarbij ik suggereer dat er 8 studies zijn gedaan naar dezelfde onderzoeksvraag, brengt me bij het volgende onderwerp. Namelijk, kunnen we meerdere studies ook samennemen om te bepalen of er wel of geen effect is? Ja, dat kan inderdaad heel goed, bijvoorbeeld met een zogenaamde meta-analyse. We kunnen daarmee het gemiddelde geschatte effect eenvoudig uitrekenen en ook wat de daarbij passende onzekerheid is. Die zal behoorlijk zijn afgenomen, omdat nu meer informatie is gebruikt. Een samenvatting van de 8 studies heb ik onder de figuur gezet.

Zoals u ziet is het gemiddelde effect over 8 studies heel klein, namelijk 0.04. En hoewel de onzekerheid is afgenomen, de breedte van het interval is van 0.9 naar ongeveer 0.3 gedaald, is dit een voorbeeld van een niet significant resultaat. We zullen dus concluderen dat er onvoldoende evidentie is voor verschillen tussen jongen en meisjes.

Het systematisch samen nemen van meerdere studies is heel belangrijk om met meer zekerheid conclusies te kunnen trekken. Er kan echter bij het samen nemen van studies van alles verkeerd gaan. Zo hebben we het probleem van publicatie bias. Zowel de onderzoekers zelf, als de tijdschriften die bepalen welke artikelen wel en niet gepubliceerd worden, geven de voorkeur aan studies waar wel effecten worden gevonden. De samenvatting overschat natuurlijk het effect als een selectief deel van de studies niet gepubliceerd en wellicht daardoor niet gevonden wordt, zoals in de tweede samenvatting te zien is. Hier ben ik er van uit gegaan dat alleen de twee significante studies gepubliceerd en dus samengenomen zijn.

De getallen zijn niet alleen verschillend; ook de conclusies over het aan- of afwezig zijn van het effect en de bijbehorende onzekerheden lopen ver uiteen. Dit geeft aan hoe ernstig we tot verkeerde conclusies kunnen komen als niet significante resultaten minder goed zichtbaar zijn dan de significante.

Daarnaast vermoeden we ook dat er *binnen* gepubliceerde studies vaak sprake is van het overschatten van het effect. Er zijn verschillende onderzoeken gedaan naar dit vermoeden en vele indicaties gevonden dat dit soort overschattingen inderdaad voorkomen. Ze zijn het gevolg van wat tegenwoordig bekend staat als questionable research practices. Deze term verwijst niet naar bewust en doelgericht frauderen, wat hopelijk maar heel zeldzaam voorkomt, maar het verwijst naar de onvermijdelijke subjectieve stappen in het analyse proces die, onbedoeld, kunnen leiden tot overschattingen van effecten. Voorbeelden zijn het bijverzamelen van data als je wel een effect in een bepaalde richting ziet maar het nog net niet statistisch significant is, of het doen van vele statistische toetsen maar in de rapportage je beperken tot de toetsen die significante resultaten hadden, alsof dat de enige uitgevoerde toetsen waren.

Als door dit soort handelingen in elke of enkele van deze studies het effect iets overschat wordt, dan zal de samenvatting natuurlijk ook een overschatting geven.

Dat de voorbeelden die ik zojuist noemde, voorbeelden zijn van subjectieve keuzes in het onderzoeksproces die niet door de beugel kunnen zou bij elke wetenschapper bekend moeten zijn. Maar zoals gezegd suggereert onderzoek hiernaar dat het wel degelijk voorkomt. Daarnaast is er ook een grijs gebied van subjectieve keuzes in het onderzoeksproces waar de lijn tussen goed en fout helemaal niet zo duidelijk is. Voorbeelden hiervan heb ik al eerder genoemd: wat doe je met extreme waarden, ofwel uitbijters, hoe ga je om met missende waarden en hoe controleer je de fouten marges als je binnen een onderzoek heel veel toetsen uitvoert?

De laatste jaren is er veel aandacht voor dit grijze gebied, evenals voor procedures om veelgemaakte fouten te voorkomen.

Questionable research practices moeten vervangen worden door responsible research practices. Wat die responsible research practices dan precies zijn is echter nog niet uitgekristalliseerd en zal nog aandacht, onderzoek en discussie vereisen. Initiatieven die al gestart zijn bevinden zich vooral op het gebied van replicatie, en op het gebied van openheid omtrent data, en de verantwoording van alle stappen in het onderzoeksproces.

Bij replicatie onderzoek ontstaat al vrij snel ook de vraag hoe je verschillende replicatie studies dan het beste kunt samen nemen om ook echt de meerwaarde van het repliceren te bereiken? Als we strak gecontroleerde studies, zoals een gerandomiseerd experiment, exact repliceren, zijn er al verschillende aanpakken voorhanden. Deze variëren van simpel en niet optimaal, zoals het tellen van het aandeel replicaties met een p-waarde onder de 5%, tot meer adequate methoden zoals bijvoorbeeld meta analyse of Bayesiaans updaten.

Vaak zullen we ons echter in een situatie bevinden waar niet meerdere exacte replicatie studies beschikbaar zijn. Er ontstaat een nieuwe uitdaging als er wel studies zijn die tot op zekere hoogte dezelfde fenomenen bestuderen. Ook hier zou je de verschillende studies willen samenvoegen om tot betere conclusies te komen. En ik ga uit van betere conclusies omdat ze gebaseerd zijn op meer informatie dan een enkele studie in isolatie geanalyseerd.

Het samen nemen van studies die behoorlijk verschillend zijn vereist echter nog meer subjectieve beslissingen en keuzes in een grijs gebied waar geen eenvoudig goed of fout bestaat. Enkele kritische reviewers van werk op dit gebied klagen dat we misschien te ver van ons ideaal van objectiviteit verwijderd raken. Toch zie ik veel potentieel.

Laat ik een voorbeeld uit de praktijk noemen. Ik werk samen met onderzoekers bij diergeneeskunde. Zij willen graag weten of bepaalde voedingssupplementen voor oudere paarden die pijnlijke gewrichten hebben tot vermindering van klachten leiden. De mate van last wordt gemeten door te kijken naar de manier van lopen van de paarden.

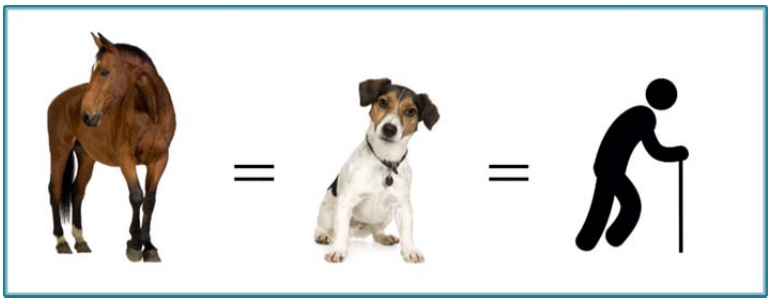
In de literatuur is een kleine eerdere studie gevonden die rapporteert dat er een significant effect gevonden is. Om dit te repliceren is in Utrecht een gerandomiseerde studie bij 24 paarden gedaan, 12 in de experimentele en 12 in de controle groep. Dat is geen grote steekproef maar de middelen zijn beperkt; onderzoek bij paarden is duur en in de diergeneeskunde gaat minder geld om dan in de humane geneeskunde. In de Utrechtse studie werd geen significant effect van de voedingssupplementen gevonden.

En nu? We hebben dus een kleine steekproef paarden waar het effect wel is aangetoond en een kleine steekproef paarden waar het effect niet werd gevonden. De studies zijn geen exacte replicaties, dus een verklaring die zeker overwogen moet worden is of het verschil in resultaten veroorzaakt kan zijn door bijvoorbeeld het verschil in hoe er gemeten is of er een effect is. Maar herinnert u zich ook nog mijn voorbeeld aan het begin? Steekproeven uit dezelfde populatie kunnen heel goed verschillende resultaten laten zien, simpelweg door toeval, en dit zal bij kleine steekproeven zelfs vaak het geval zijn. Er rest ons dus niks anders dan te concluderen dat we het nog niet weten en dat er echt een grotere –wellicht onbetaalbare- paardenstudie nodig is om conclusies te kunnen trekken. Of toch niet?

Een gesprek met een expert op het gebied van dit soort gewrichtsproblemen bij paarden leert ons dat de volgende redenering in de diergeneeskundige praktijk niet ongewoon is.

Er zijn ook onderzoeken gedaan bij honden met dit soort problemen. Laten we eens kijken of daar effecten gevonden zijn. En hoewel we natuurlijk niet kunnen spreken van een exacte replicatie, een hond is geen paard, zijn er wel degelijk grote overeenkomsten tussen paarden en honden in hoe de gewrichten werken en hoe ouderdomsproblemen hierbij een rol spelen.

Sterker nog, onze expert geeft aan ook waarde te hechten aan studies naar dit soort voedingssupplementen bij mensen met vergelijkbare problemen.



Wat er in de praktijk nu zou kunnen gebeuren is dat een dierenarts op grond van de significante paardenstudie en een bij hem of haar bekende hondenstudie die ook effectiviteit liet zien het middel voortaan bij paarden zal toedienen. En een andere dierenarts die de Utrechtse niet significante studie kent en daarnaast net de resultaten van een grote humane studie heeft gelezen waarin eveneens geen effect werd aangetoond zal wellicht op zoek gaan naar een andere behandeling of medicatie.

Wie heeft er gelijk? Er zal geen discussie zijn over het feit dat er nog onzekerheid is. Maar als we de beschikbare plukjes evidentie zo laten als het is, kan elke dierenarts zijn eigen betoog bouwen met een subjectief gekozen deel van die evidentie en zelfs beargumenteren dat zijn handelen evidence based is. Wat ik bepleit is het formeel synthetiseren van de verschillende bronnen van informatie, waarbij een expert op dit vlak mag aangeven hoe relevant de informatie is voor de onderzoeksvraag die centraal staat. Hierin kunnen allerlei elementen worden meegenomen: verschillen in de behandeling, verschillen in hoe de effectiviteit is gemeten, en verschillen in diersoort.

Het moge duidelijk zijn dat dit niet eenvoudig is. Het vereist veel domeinkennis van de expert en verschillende experts zullen niet exact dezelfde mening hebben. Veelal zullen we meerdere experts raadplegen om variatie in meningen hierover in kaart te brengen, of ze bijvoorbeeld middels een Delphi studie stimuleren om consensus te bereiken. Ook zullen we altijd een zogenaamde sensitiviteitsanalyse doen. Dit wil zeggen dat we de analyse meerdere keren uitvoeren met bijvoorbeeld de verschillende meningen van experts als input. Op deze wijze kun je zien of je conclusies relevant verschillen als je andere aannames maakt over de relevantie van andere studies in het veld.

We hebben het hier niet over een wondermiddel. Het feit dat er nog veel onzekerheid is na twee kleine paardenstudies die verschillende resultaten toonden is onveranderd. Indien mogelijk is een nieuwe, grotere paardenstudie de beste manier om meer zekerheid te krijgen. En dan liefst ook nog een paar replicatiestudies, want ook grotere steekproeven hebben nog last van toeval.

Maar wanneer dit niet haalbaar is, of nog niet, zal het vaak voorkomen dat menselijk redeneren leidt tot het op niet doorzichtige of controleerbare wijze synthetiseren van beschikbare informatie. Ik pleit voor methoden die dit proces transparant maken. Allerlei subjectieve keuzes en meningen zullen invloed hebben op welke kennis en eerdere studies wel en niet worden meegenomen. Maar door deze stappen onderdeel van de methoden en analyse van je huidige studie te maken, forceer je op argumenten gebaseerde keuzes en een controleerbaar proces.

Deze ideeën passen ook heel goed in een andere uitdaging waar onderzoekers en methodologen tegenwoordig mee te maken hebben, namelijk de enorme toename van data.

Ik heb het eerder vandaag gehad over de beperkte waarde van steekproeven die aan de kleine kant zijn. Misschien herinnert u zich nog dat ik het voorbeeld over ruimtelijk inzicht was gestart met te zeggen "laten we er van uit gaan dat de data op een optimale wijze is verzameld en alles zonder systematische fouten is gemeten". In dat geval is het resterende probleem niet meer dan de onzekerheid die je hebt bij het vertalen van je resultaten in de steekproef naar conclusies over de populatie waaruit de steekproef getrokken is.

Met een zeer simpel voorbeeld kon ik eenvoudig illustreren wat de rol van steekproef-groottes op de onzekerheid van bevindingen is, hoe p-waarden hier erg gevoelig voor zijn en dat het daarom van belang is alternatieve maten en methoden te overwegen.

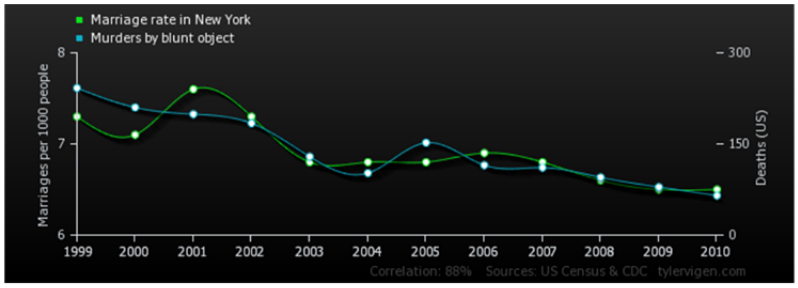
Bij het iets minder simpele voorbeeld uit de diergeneeskunde, heb ik verteld over het samennemen van studies die niet identiek waren in opzet en uitvoering, of zelfs verschilden in het soort dier dat centraal stond.

Echter, in beide voorbeelden ging ik er nog wel vanuit dat de onderzoeker bepaalt welke data verzameld wordt om de onderzoeksvraag te beantwoorden. Tegenwoordig hebben we ook steeds meer zogenaamde 'big data', dat wil zeggen, een heleboel informatie die niet doelbewust verzameld is maar die gewoonweg beschikbaar zijn.

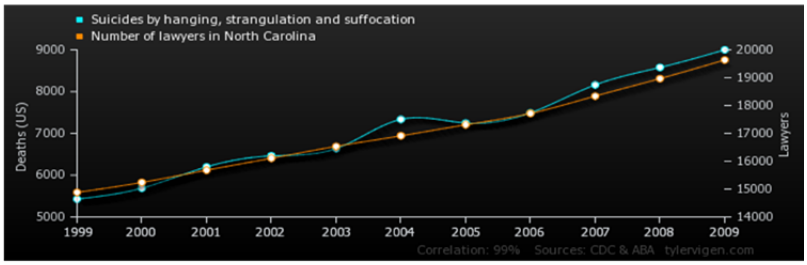
Bekende voorbeelden hiervan zijn sensor data, GPS data, twitter data, of bijvoorbeeld gebruikte google zoektermen. Laatst genoemde heeft geleid tot het meest bekende voorbeeld van zowel de kracht als het risico van het maken van voorspellingen op grond van big data. Google publiceerde in 2009 trots dat ze sneller dan de gezondheidsinstanties kon voorspellen wanneer en waar er een griepepidemie was, op grond van de gebruikte zoektermen van internetgebruikers.

U kunt zich voorstellen dat tegenwoordig veel mensen, voordat ze naar een arts gaan, al gegoogeld hebben op allerlei griep-gerelateerde termen. Door het registreren en analyseren van deze zoektermen kon Google dus eerder dan de officiële instanties signalen van een toename in griepgevallen oppikken.

Een aantal jaren later zat Google er met hetzelfde voorspellingsmodel echter volledig naast. Een mogelijke verklaring hiervoor is dat in big data sets naar relaties tussen verschillende aspecten wordt gezocht, zonder dat er genuanceerd en theorie-gestuurd wordt nagedacht over oorzaak en gevolg en over mogelijke alternatieve verklaringen voor de gevonden statistische relaties. Het is heel goed mogelijk dat twee variabelen sterk correleren zonder dat ze betekenisvol samenhangen. Het internet staat vol met voorbeelden van 'grappige maar niet betekenisvolle correlaties'.



Correlation: 0.879854



Correlation: 0.993796

Zo blijkt er een sterk verband te zijn tussen het aantal huwelijken dat in een jaar in New York gesloten wordt en het aantal moorden met een stomp voorwerp in dat zelfde jaar.

In het tweede voorbeeld zien we een nog sterker verband tussen het aantal zelfmoorden door ophanging of verstikking en het aantal advocaten dat North Carolina heeft.

Deze voorbeelden heb ik gevonden op een website die opent met de zin: "elke dag een nieuwe spurieuze correlatie". Dit laat twee dingen zien. Ten eerste dat er ongelooflijk veel informatie wordt geregistreerd en meer of minder openbaar beschikbaar is om te analyseren. En ten tweede dat deze enorme hoeveelheden data gebruikt kunnen worden om interessante of totaal oninteressante correlaties op te sporen.

Voor wetenschappers, waaronder onderzoekers in de sociale wetenschappen, en methodologen en statistici liggen hier kansen en uitdagingen. Het werken met extreem grote data sets evenals het nadenken over wat je wel en niet kunt leren van dit type data wordt wel aangeduid met de term 'data science'. Maar data science kan breder geïnterpreteerd worden dan alleen het big data gedeelte. Onder data science verstaan we ook grote onderzoeksprojecten waar data van verschillende aard wordt verzameld door een interdisciplinair team van onderzoekers. Denk hierbij bijvoorbeeld aan een mix van kwalitatieve data, data uit experimenteel onderzoek, en FMRI data.

De middelen om grote hoeveelheden informatie te verzamelen en op te slaan zijn in de afgelopen jaren in hoog tempo toegenomen. Denk aan sensoren die mensen of dieren op of in het lichaam dragen, zodat continue allerlei metingen worden gedaan. Denk aan GPS tracking en allerlei apps die we op onze mobieltjes geïnstalleerd hebben die beweging en verplaatsingen registreren. Maar denk ook aan medische ontwikkelingen op het gebied van bijvoorbeeld genetisch materiaal of hersenscans.

Veel minder verandering is er in de doelen van ons onderzoek. We willen gedrag van mensen, instituties, of samenlevingen kunnen voorspellen en begrijpen. We bevragen nog steeds mensen middels vragenlijsten en we observeren nog steeds zowel in natuurlijke settings als in gecontroleerde experimenten. We staan echter op het punt dat we deze zorgvuldig verzamelde data graag willen integreren met overige beschikbare databronnen om zo tot beter begrip of betere voorspellingen te komen.

Om deze integratie mogelijk te maken is kennis nodig over de verschillende vormen van data en de daarbij passende analyse technieken. Om goed te integreren is ook kennis nodig over de bestudeerde fenomenen; de zogenaamde domein kennis. Ook hier hebben we te maken met een mix van expertises, bijvoorbeeld kennis vanuit de ontwikkelingspsychologie en neuropsychologie maar ook medische kennis.

Kortom, synthese staat ook hier centraal. Synthese van verschillende vormen van data, verschillende analyse technieken en domein specifieke kennis uit meerdere domeinen. Dit vraagt om vergaande samenwerkingen tussen de disciplines en zal leiden tot nieuwe methoden en inzichten. Ik verwacht echter niet dat er kant-en-klare recepten zullen ontstaan voor de aanpak van dit soort interdisciplinaire vraagstukken waarbij gevarieerde bronnen van data gebruikt worden. Er zullen steeds weer afwegingen en keuzes gemaakt moeten worden, waarbij domein specifieke kennis naar mijn idee een cruciale rol zal hebben. Ook hier zal volledige objectiviteit nooit haalbaar zijn. Dat moeten we wat mij betreft maar zo vroeg mogelijk erkennen, omdat dat het transparant rapporteren over de subjectieve keuzes die gemaakt moeten worden, zal bevorderen.

En hiermee ben ik wederom bij de titel van deze rede gekomen, en naderen we het einde. Graag vat ik kort samen hoe de door mij besproken onderwerpen zich verhouden tot een aantal van mijn concrete onderzoeksactiviteiten. Als dat achter de rug is resten nog drie losse eindjes: excuses, vrouwen in de wetenschap, en een woord van dank.

Ik heb het gehad over integere en transparante wetenschap waarin subjectieve keuzes onvermijdelijk zijn. In dit kader verdienen de huidige, wellicht te rigide toegepaste, statistische procedures een kritische blik. Is het formuleren en wel of niet verwerpen van een nulhypothese wel passend, of best-passend, bij de onderzoeksvragen die we hebben? En rapporteren en interpreteren we wel duidelijk genoeg de onzekerheden bij onze schattingen of toets resultaten? Dit thema is onlosmakelijk verbonden met mijn onderzoek naar het gebruik van Bayesiaanse methoden in het algemeen en het evalueren van informatieve hypothesen in het bijzonder.

Daarnaast heb ik het gehad over het synthetiseren van onderzoek. Hoe kunnen we op transparante wijze eerder gedaan onderzoek meenemen in de analyse van nieuw onderzoek. Aan dit tweede onderzoeksthema werk ik vooral samen met zeer gewaardeerde collega's van epidemiologie en diergeneeskunde. Centraal staat wederom een goede inschatting van de effecten en de daarbij behorende onzekerheden. Welke externe of eerdere informatie je wel of niet meeneemt, evenals op welke wijze je dit includeert, vereist subjectieve keuzes. Dit geldt voor meer traditionele methoden, zoals de meta analyse, en nog sterker voor Bayesiaanse analyses met informatieve priors.

Hieraan gerelateerd, maar met een wat andere invalshoek, is het zoeken naar integratiemethoden, wanneer studies en bijbehorende datasets van heel verschillende aard zijn. Hoe kunnen we zinvol kwalitatieve data koppelen aan kwantitatieve data? Hoe kunnen we zinvol de resultaten van meer exploratieve analyses op bijvoorbeeld FMRI of genetische data koppelen aan experimentele data en daarbij ook nog theoretische kennis van de verschillende domeinen meenemen? Dit zijn uitdagingen waar ik hopelijk de komende jaren veel over na mag denken. Het is ook het soort project dat alleen succesvol kan worden in een samenwerking met onderzoekers uit verschillende disciplines met elk hun eigen theoretische kennis en bijdragen, evenals met collega methodologen met een verscheidenheid aan expertises. We zijn nog maar net gestart op dit gebied, en het heeft nu nog vooral de vorm van oriënteren op het probleem, brainstormen over aanpak en doelen, en een eerste pilot project. Ik verheug me op al onze toekomstige ideeën en zal mijn uiterste best doen ervoor te zorgen dat ze uitgevoerd worden.

En dan is het tijd voor excuses.

Ondanks dat 45 minuten heel erg lang leek toen ik begon met schrijven, bleek er toch niet voldoende tijd te zijn om alle onderwerpen die me aan het hart gaan de verdiende aandacht te geven. Mijn eerste excuses gaan uit naar het onderwijs en alle collega's die zich daar zo hard voor inzetten. Mijn passie voor onderwijs is ongewijzigd en toch is deze kant van mijn werk vandaag wat onderbelicht. En dat terwijl mijn persoonlijke omgeving zal beamen dat ik niet vrolijker en enthousiaster thuis kan komen dan na een dag waarop ik een inspirerende interactie heb gehad met studenten, of het nu een college voor onze research master studenten was, of een overleg met studenten die bijvoorbeeld aan hun bachelor- of masterthesis werken.

De tweede verontschuldiging heeft te maken met onderzoek. Naast alles waarover ik vandaag heb gesproken loopt er ook nog een vidi project op het gebied van circulaire data en de statistiek daarvoor. Ook daar had ik best 45 minuten mee kunnen vullen, hoewel ik niet zeker weet of iedereen dan nu nog wakker zou zijn. Maar dat is misschien sowieso ijdele hoop. Jolien en Kees: jullie zijn het circulaire data project. Door jullie werk en enorme talent en onafhankelijkheid loopt het project goed; door jullie enthousiasme en persoonlijkheden is het ook echt leuk. Sorry dat het vandaag in het geheel niet over circulaire statistiek ging. Ik hoop dat jullie hebben opgemerkt dat ik, om dit gemis enigszins te compenseren, hier en daar wat cirkels in de slides heb verwerkt.

Voor mij persoonlijk is dit natuurlijk een belangrijke dag. Het is de bezegeling van het bericht dat ik in oktober 2015 van onze decaan professor Werner Raub ontving, waarin stond dat ik was benoemd als hoogleraar. Daar is toen al champagne op gedronken.

Vandaag, op 10 februari 2017, heb ik de gelegenheid gekregen om mijn beoogde invulling aan dit hoogleraarschap toe te lichten. Om die reden zal 10 februari 2017 voor mij altijd een bijzondere datum blijven.

Maar vandaag is ook een historische dag. Het is exact 100 jaar geleden dat de eerste vrouwelijke hoogleraar van Nederland is benoemd, en dat was hier aan de Universiteit van Utrecht. Op 10 februari 1917 om 2 uur 's middags sprak Johanna Westerdijk haar oratie uit. Vandaag om 2 uur opende minister Jet Bussemaker hier in het gebouw de startbijeenkomst van de Westerdijkviering.

Ik vind het schokkend dat het feitelijk nog maar zo kort geleden is dat het voor vrouwen mogelijk was om een carrière in de wetenschap te maken. En hoewel mogelijk, was het verre van makkelijk. Dat blijkt wel uit het feit dat het na Johanna Westerdijk nog eens 30 jaar heeft geduurd voordat de tweede vrouwelijk hoogleraar in Nederland werd benoemd.

Dat de weg naar het hoogleraarschap voor Johanna Westerdijk geen geplaveide route was zal duidelijk zijn. Zij wordt wel beschreven als een sterke persoonlijkheid, een beetje opstandig, met lak aan conventies en zich veelal gedragend als 'one of the boys'. Ook 100 jaar later zullen maar weinig hoogleraren de route naar het hoogleraarschap als een geplaveid pad beschrijven en helaas is het denk ik ook nog steeds zo dat er tussen mannen en vrouwen verschillen zijn in hoe moeilijk of makkelijk verschillende carrière-stappen te bereiken zijn. Desalniettemin ben ik dankbaar dat ik hier vandaag mag staan, zonder mezelf geweld aan te hebben hoeven doen en zonder me genoodzaakt te voelen me als 'one of the boys' te gedragen. We zijn er nog lang niet maar er is zeker wel wat bereikt in die 100 jaar.

En dan tot slot een woord van dank.

Allereerst wil ik het college van bestuur en het bestuur van de faculteit sociale wetenschappen bedanken voor het in mij gestelde vertrouwen. Tevens wil ik de Universiteit van Twente danken, die een jaar eerder mijn benoeming als bijzonder hoogleraar bij de afdeling OMD heeft mogelijk gemaakt. Hierbij noem ik graag expliciet professor Cees Glas, toenmalig afdelingsvoorzitter en professor Bernard Veldkamp, huidig afdelingsvoorzitter, die mij verwelkomd hebben in hun groep en beiden vandaag aanwezig zijn.

Op het meer persoonlijke vlak, kan en wil ik niet anders dan beginnen met een woord aan Herbert. Ik ben begonnen als docent statistiek maar wist al snel dat ik ook graag wilde promoveren. Vanaf het moment dat ik die kans kreeg, met dank aan toenmalig afdelingsvoorzitter Peter van der Heijden en het emancipatiefonds waar financiering werd gevonden, heb ik vooral van Herbert geleerd. Dit betrof zowel kennis over Bayesiaanse statistiek, programmeren en artikelen schrijven, als zaken omtrent wetenschappelijke waarden en leiderschap. Volgens mij had ik geen betere start van mijn wetenschappelijke carrière kunnen hebben. Als ik dan tot slot hier aan toevoeg dat ik je ook persoonlijk graag mag, wil ik daar je wederhelft Anja zeker ook expliciet bij noemen.

Daarnaast is de hele afdeling methodenleer en statistiek me heel dierbaar. Toen ik als 'jonkie' bij M&S begon, waardeerde ik vooral de manier waarop de meer ervaren docenten en onderzoekers de jongere generatie stimuleerde en hoe men bereid was ons te onderwijzen en te coachen, en ons de gelegenheid gaven om te leren, uit te proberen, te struikelen en te groeien. M&S was voor mij zowel een inspirerende als een veilige werkomgeving en ik ben er van overtuigd dat dit van groot belang is voor zowel het geluk als het presteren van alle leden van de groep.

Nu ik zelf niet meer kan ontkennen dat ik aan de seniore kant sta, gaat mijn waardering juist heel erg uit naar het enorme enthousiasme van de jongere collega's. Wat een fijne mensen zijn al onze promovendi en junior docenten. Zeer getalenteerd en gepassioneerd geven zij een onmisbare boost aan de werksfeer op de afdeling. Dat ze daarnaast ook voor de nodige borrels en feesten zorgen is de kers op de taart.

Eén collega wil ik toch nog bij naam noemen. Leoniek, ik denk dat je wel weet hoe belangrijk onze gesprekken voor mij zijn. Wanneer nodig geef jij m'n zelfvertrouwen een zet en moedig je me aan met zowel concrete adviezen als morele en emotionele support. Hoewel we elkaar maar weinig buiten het werk ontmoeten ben jij veel meer vriendin dan collega. Maar ik ben boven alles blij dat we beiden zijn, want ook als collega zou ik je niet graag missen, om al het goede werk dat je doet voor zowel de afdeling als voor de universiteit.

En natuurlijk is er nog een heel leven buiten de wetenschap. Hierin geniet ik van het gezelschap en de liefde van mijn familie, schoonfamilie, vrienden en vriendinnen, sportmaatjes, en reisgenoten. Qua namen noemen beperk ik me tot twee.

Hans en Anouk: het leven is een stuk leuker, gezelliger, en liefdevoller geworden sinds ik jullie ken. Ook wat chaotischer en veel ingewikkelder maar het is het zeker waard.

Anouk: je bent een prachtige meid, van buiten en van binnen, en ik ben hartstikke trots op je!

Hans, morgen hebben we elkaar 3 jaar geleden voor het eerst ontmoet. De roze bril is inmiddels toch verdwenen; maar de liefde niet. Ik geniet nog elke dag van ons.

Ik heb gezegd.