

Bachelor Thesis

Impact van de gebruikte statistische methode op een kleine dataset

Inge van der Borg (5768845)

Universiteit Utrecht

Thesis Pedagogische Wetenschappen

200600042

Begeleider

Rens van de Schoot, PhD.

Begeleider

Sanne C. Smid, MSc.

2^e beoordelaar

prof. dr. I.G. Klugkist

Datum van inleveren:

23 juni 2017

Abstract

As a researcher you have to decide in advance what statistical method you use to analyse your data. When dealing with small sample sizes, decisions in statistical method can have a major impact on the results. In this thesis six exact datasets were generated to investigate the differences in results. Each dataset consisted of four groups. The datasets differ in effect size (small/medium/large) and power (0,4 & 0,8) and were analysed using: ANOVA, Contrast test, F-bar test, Non parametric (K&W-test), bootstrap and Bayesian statistics. The results showed that the method used could be of great impact on the final conclusion. Especially the methods where an informed hypothesis was tested, stood out in a positive way. Based on this study we recommend that if you have information about your data in advance, you should try to implement it in the chosen statistical method before testing.

Keywords: Data, Sample size, ANOVA, Non-parametric, Bayesian statistics

Impact van de gebruikte statistische methode op een kleine dataset

Om uit te komen in de juiste haven, wordt de koers over de grond bepaald. Een halve graad afwijking van deze koers zou het verschil kunnen maken of je in bijvoorbeeld New York of Philadelphia uitkomt. Welke haven moet het schip uitkomen? Welke koers over de grond is hiervoor nodig? Vragen die een navigatieofficier aan boord van een schip zich zeker gesteld heeft vóór hij zijn reis ging maken.

Ook een wetenschappelijke onderzoeker moet van tevoren veel beslissingen maken over de 'koers' die gevaren gaat worden. Zo moet onder andere het significantielevel worden vastgesteld en de manier van dataverzameling. Daarnaast is de statistische methode die gekozen wordt om de data te analyseren van belang. De statistische methode zou het verschil kunnen maken tussen het verwerpen of niet verwerpen van de nulhypothese. Zou de keuze van statistische methoden daadwerkelijk het verschil kunnen maken bij een kleine steekproef?

Een grote uitdaging binnen het doen van onderzoek is het verzamelen van voldoende relevante data. Daarbij zijn situaties denkbaar waarin het niet mogelijk is voldoende data te verzamelen in een populatie en zal gewerkt moeten worden met een kleine steekproef. In veel literatuur over statistische modellen worden uitspraken gedaan over de steekproefomvang die nodig is om verschillen tussen de groepen weer te geven wanneer er ook verschil aanwezig is. Er kan dan gezocht worden naar een methode waarbij dit verschil ook gevonden wordt en het model significant is, ook met een kleine dataset. Hierbij zijn de effectgrootte, het significantielevel (α) en de power ($1-\beta$) waarmee gerekend wordt van belang (Field, 2013).

Power

De power is de kans dat de statistische test een effect vindt wanneer er een effect bestaat. Power is belangrijk omdat dit wat zegt over de kans dat een type I fout gemaakt wordt. Wanneer er een effect wordt gevonden die in werkelijkheid niet aanwezig is, heet dat

een type I fout. Andersom noemen we de kans dat je de nulhypothese onterecht verwierpt omdat er geen effect wordt gevonden maar er in werkelijkheid wel is, een type II fout. De power wordt vooraf berekend aan de hand van de verwachte effectgrootte om zo de steekproefgrootte te bepalen en de kans op een type I fout. Hiermee kan bepaald worden of een onderzoek waarschijnlijk succesvol zal zijn of niet. In figuur 1 (Magnusson, z.j.) is dit grafisch weergegeven. De power kan alleen berekend worden wanneer de distributie normaal is, aangezien bij een andere verdeling er niet vanuit gegaan kan worden dat de power samenhangt met de type I fout (Field, 2013). Wanneer de power echter wordt berekend uit een steekproef die normaal verdeeld is, zal blijken dat de non-parametrische test minder power heeft dan de parametrische test (Field, 2013). Het gevaar voor kleine steekproeven is dat wanneer zij niet voldoen aan de centrale limietstelling, ze niet beschouwd kunnen worden als normaal verdeeld, waardoor aan deze aanname niet kan worden voldaan en de power niet berekend kan worden. Voldoende Power is belangrijk omdat dit de kans op een type I fout aanzienlijk verkleint. Zeker binnen een kleine dataset, omdat de invloed van steekproeffouten dan groter is.

Significantielevel alfa

In de sociale wetenschap wordt vaak gebruikgemaakt van een alfalevel van 0,05 dit betekent dat er 5 % kans is dat het gevonden effect in werkelijkheid niet bestaat wanneer dit wel wordt aangenomen (Type I fout). Dit wordt algemeen geaccepteerd door wetenschappers als 'goed genoeg' (Field, 2013).

Effectgrootte

Alleen significantie vertelt nog niks over het effect. Hoe groter het effect is, hoe makkelijker dit effect ook gevonden zal worden in een steekproef. Om de grootte van het effect weergegeven wordt bij een ANOVA gebruikgemaakt van de Cohen's *f*. Dit is een

gestandaardiseerde maat dat iets zegt over de sterkte van het geobserveerde effect (Field, 2013).

Huidige thesis

Als onderzoeker kies je van te voren een statistische methode om je data te analyseren. Op dat moment weet je nog niet wat de verschillen zijn tussen de resultaten van de verschillende methodes en welke impact deze beslissing kan hebben op de eindconclusie. Mede doordat Power, effectgrootte en steekproefgrootte ook van invloed zijn op de resultaten. Om deze impact zichtbaar te maken wordt in deze thesis een model gebruikt met vier groepen. Door dit model te testen met verschillende statistische methodes kan worden gekeken of er statistische significantie bereikt kan worden. De onderzoeksvraag die wordt gesteld: hangt de beslissing van de onderzoeker over statistische significantie af van de methode die gebruikt zal worden bij een kleine dataset?

Om de onderzoeksvraag te beantwoorden wordt exacte data gegenereerd met verschillende effectgroottes en power. Deze datasets bestaan zoals gezegd uit 4 groepen; μ_1 , μ_2 , μ_3 en μ_4 . Deze datasets worden getest met verschillende statistische methodes. Hieronder volgt een korte uitleg van de gebruikte statistische methodes in deze thesis. Vervolgens zullen de specificaties van de datasets worden gegeven en hoe deze zijn geanalyseerd. In de resultatensectie zijn de resultaten van de analyses zichtbaar. Tot slot worden de resultaten besproken in de discussie. De limitaties en aanbevelingen vanuit deze thesis worden hier als laatste genoemd.

Theorie statistische methodes

In de volgende sectie zullen de gebruikte statistische methodes kort worden uitgelegd.

ANOVA

De ANOVA is een model die gebruikt wordt om hypothesen te testen. In dit voorbeeld wordt de hypothese $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ getest. Hierbij worden de gemiddelde tussen twee of

meer groepen vergeleken. Bij de ANOVA wordt de F -ratio gebruikt om dit te kunnen testen. Dit model gaat uit van een paar aannames, namelijk: Er moet een aselechte steekproef getrokken zijn. De observaties moeten onafhankelijk zijn. De afhankelijke variabele moet in de populaties normaal verdeeld zijn en de populaties moeten gelijke variantie hebben (homoscedasticiteit) (Gravetter & Wallnau, 2016). Bij een kleine steekproef kan je niet altijd vertrouwen op de centrale limietstelling. Hierdoor zal er niet worden voldaan aan de aanname van de normale verdeling. Wanneer deze aanname niet gehaald kan worden, zou ook een ‘non-parametrische test’ kunnen worden uitgevoerd (Field, 2013).

Non-parametrisch testen

Zoals al eerder beschreven, kan non-parametrisch testen een uitkomst bieden wanneer gewerkt wordt met kleine steekproeven. Dit komt omdat bij deze testen niet voldaan hoeft te worden aan de aannames van parametrische statistiek. Bij Non-parametrische testen wordt een rangorde toegekend aan de originele scores uit de steekproef. Hierbij wordt dezelfde hypothese getoetst als bij de ANOVA. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. Met deze statistische methode kunnen vreemde verdelingen toch geanalyseerd worden (Field, 2013).

Bootstrap test

Een andere optie is de bootstrap test. Een bootstrap test heeft als voordeel boven een gewone ANOVA, dat er niet aan de aanname van een normaalverdeling voldaan hoeft te worden. De bootstrapmethode ondervangt dit door de steekproef te zien als de gehele populatie en heel veel mogelijke kleinere steekproeven te trekken uit deze ‘populatie’. SPSS heeft als default setting dat er 1000 steekproeven worden getrokken uit de originele steekproef om zo tot het 95% bootstrap betrouwbaarheids interval te komen. Dit betekent dat 95% van de bootstrap steekproeven valt binnen deze parameters. Nu kunnen uit de bootstrap samples het gemiddelde en de standaarddeviatie worden bepaald (Field, 2013). Ook hier wordt de nul hypothese getoetst $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

Contrast test

Wanneer een ANOVA wordt uitgevoerd kan aangegeven worden of er een verschil is tussen de groepen en of dat verschil significant is. Er kan echter na aanleiding van de ANOVA niet geconcludeerd worden waar dit verschil tussen groepen zich bevindt. Hiervoor zijn aanvullende tests nodig. Deze kunnen gedaan worden door post hoc tests of door contrast testen. Bij de laatste ga je ervan uit dat je al gericht een hypothese kan opstellen. Door een bepaald gewicht toe te kennen aan de groepen kan er gericht worden getoetst. Hierbij geef je waarde aan de te testen hypothese. In ons voorbeeld: $\mu_4 > \mu_3 > \mu_2 > \mu_1$. Door nu gewicht te geven aan de groepen, bijvoorbeeld: $\mu_4 = 1,5$, $\mu_3 = 0,5$, $\mu_2 = -0,5$, $\mu_1 = -1,5$, Kan je de voorspellende waarde toekennen zoals in de hypothese is geformuleerd. Een contrast test bestaat uit verschillende stappen. Je kan telkens maar 2 groepen vergelijken. Het begin is dus: Een referentiegroep met de overige groepen. Vervolgens wordt deze laatste groep weer opgesplitst in een referentiegroep en de overige groepen, net zolang tot alle groepen (k) aan bod zijn geweest (k-1). In elke contrast test is een bepaald gewicht bepaald dat aangeeft hoe de hypothese was opgebouwd.

F-Bar test

Bij een F-bar test wordt ook gebruik gemaakt van een informatieve hypothese zoals $H_i: \mu_4 > \mu_3 > \mu_2 > \mu_1$, hierbij worden geen afstanden tussen de gemiddeldes verondersteld zoals dat bij contrast testen wel het geval is. Dit heeft als voordeel boven nulhypothese testen dat er niet meerdere tests uitgevoerd hoeven worden omdat je direct de verwachting kan toetsen. Hierdoor heb je dus niet meer te maken met kanskapitalisatie en het geeft meer power voor de analyse (Vanbrabant, Van de Schoot, & Rosseel, ter perse). De F-bar test kan in het software programma 'restriktor' in 'R' worden uitgevoerd. Restriktor presenteert een F-bar global test. Hierbij is de nulhypothese getest ten opzichte van de informatieve hypothese, waarbij de parameters van de nulhypothese alle 0 bedragen (behalve het eventuele

intercept). En de informatieve hypothese de parameters zijn aangepast aan de ingevoerde constraints. De P-waarde geeft nu aan of de nulhypothese verworpen dan wel niet verworpen moet worden.

Bayesiaanse statistiek

Elk onderzoek dat wordt beoordeeld met frequentistische statistiek modellen kan ook worden beoordeeld door de statistische methode van Bayes. De theorie van Bayes verschilt echter op een aantal punten. Allereerst wordt binnen de frequentistische methodes gebruik gemaakt van de verdeling van de populatie en alle mogelijke steekproeven die hieruit getrokken kunnen worden. Op basis daarvan wordt de waarschijnlijkheid van jouw steekproef bepaald en of de nulhypothese wordt verworpen of juist niet. Binnen de theorie van Bayes wordt niet gebruik gemaakt van een waarschijnlijkheid van steekproef maar uitgegaan van een verwachte verdeling van de populatie. Deze verdeling wordt de ‘prior’ genoemd. Vervolgens wordt de data verzameld en deze geeft ook een verdeling. Dit is de waarschijnlijke verdeling. Deze twee verdelingen worden samen gewogen en geven de ‘posterior’ verdeling. Dit betekent dat er niet wordt gezocht naar één parameter maar dat binnen de Bayesiaanse statistiek wordt gewerkt met een waarschijnlijkheidsinterval van de parameter.

Van de Schoot et al. 2014, geven in hun artikel een mooie vergelijking tussen Bayes en frequentistische statistiek. Bij deze laatste gaat het om de kans dat een parameter voorkomt. Dit valt te berekenen door alle mogelijke steekproeven te trekken en de kans te berekenen dat jouw gevonden gemiddelde en standaarddeviatie overeenkomt met de steekproevenverdeling die gebaseerd is op alle mogelijke steekproeven. Bij Bayes echter gaat het meer om een weddenschap. Je sluit pas een weddenschap met iemand af als je overtuigd bent van je eigen gelijk, deze overtuiging haal je uit je eigen ervaringen en misschien ook wel uit voorgaande kennis. Met deze achtergrond informatie sta je zelfverzekerder in je

wedenschap. Wanneer je uitkomst echter anders blijkt pas je deze kennis weer toe op je reeds bestaande kennis.

Een prior is dus eigenlijk bestaande kennis die meegenomen wordt in de analyse om samen met een nieuwe steekproef tot een verbeterd beeld van de werkelijkheid te komen.

Bayes factor

Bij frequentistische statistiek wordt gewerkt met de nulhypothese (effect = 0). Deze nulhypothese wordt niet verworpen wanneer blijkt dat de scores uit de getrokken steekproef niet afwijkt en verworpen wanneer de scores wel afwijken. Dit gebeurt vaak door de proportie te berekenen. Dit is de waarschijnlijkheid van de getrokken steekproef ten opzichte van alle mogelijke steekproeven (Gravetter & Wallnau, 2016).

Bij Bayesiaanse statistiek wordt vaak niet gewerkt met een nulhypothese. Hier wordt bekende informatie meegenomen in de analyse. Zo kunnen ook meer gerichte hypothesen worden opgesteld, dit worden informatieve hypothesen (H_i) genoemd. In een informatieve hypothese kan al richting bepaald worden. Bijvoorbeeld: effect > 0 of effect < 0. Nu valt te berekenen of deze informatieve hypothese beter aansluit bij de data dan een andere hypothese (H_u). Deze factor van verbetering of verslechtering ten opzichte van een andere hypothese noemen we de Bayes factor.

$$Bayes\ Factor\ H_i, H_u = \frac{Fit_{H_i}}{Complexiteit_{H_i}}$$

Bayes factors in JASP

In JASP toets je de nul hypothese ten opzichte van de alternatieve hypothese een prior is bij dit programma vertegenwoordigd door een getal voor 'R Scaled fixed effects'. Dit vertegenwoordigd een verwachte effectgrootte (Rouder, Morey, Verhagen, Swagman & Wagenmakers, ter perse). Zoals al eerder gezegd is JASP nog in ontwikkeling voor

Bayesiaanse ANOVA en kan ten tijde van dit schrijven geen informatieve hypothese getoetst worden. We toetsen hier de nul hypothese $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

Bayes factors in BIEMS

In BIEMS kan wel een informatieve hypothese worden getoetst: $H_1: \mu_4 > \mu_3 > \mu_2 > \mu_1$, hierbij wordt de informatieve hypothese naar de prior getransformeerd. Het verschil met contrasttoetsen is dat er geen afstand wordt aangenomen tussen de groepen.

Het voordeel van werken met een Bayes factor is dat je niet werkt met absolute getallen. Er valt dus niet te zeggen of een bepaalde waarde hoog of juist laag is. Dit hangt af van de invulling die eraan gegeven wordt. Wel valt te zeggen dat een Bayes factor van 1 betekent dat de vergeleken hypothesen ten opzichte van elkaar niet verschillen in hoe goed ze aansluiten bij de data. Vanzelfsprekend sluit een hypothese met een Bayes factor > 1 dus beter aan dan een hypothese met een Bayes factor < 1 . Volgens Jefferys (1961) en Kass & Raftery (1995) zijn alternatieve hypothesen met een Bayes factor tussen de 1 en 3 'not worth a bare mention' - niet noemenswaardig – een Bayes factor tussen de 3 en 20 heeft een positieve verbetering ten opzichte van de nulhypothese, tussen de 30 en 150 een sterke verbetering en boven de 150 heeft de alternatieve hypothese een sterke verbetering ten opzichte van de nulhypothese. Deze getallen zijn net zo arbitrair als de frequentistische cut-off waarde van 0,05 maar samen met de Bayes factor van 3 worden ze wel gebruikt als een soort van beslissing of er wel/geen effect is. In de discussie zal dit nog worden besproken.

Methode

In dit onderzoek is gekozen voor een dataset die bestaat uit vier groepen. Voor het genereren van data is gebruik gemaakt van het programma G*power 3.1.9.2. (Erdfelder, Faul, & Buchner, Behavior Research Methods, Instruments, & Computers, 1996)

Onder de tab 'Central and noncentral distributions' zijn F tests, ANOVA: fixed effects, omnibus, one-way en A priori: Compute required sample size – given α , power, and

effect size geselecteerd. Er is uitgegaan van vier groepen met een α van ,05. Voor de effectgrootte zijn een klein, middel en groot effect ingevuld met de respectievelijke waarde voor Cohen's f van: 0,10, 0,25 en 0,40. Dit is gedaan bij een lage power van 0,40 en dan nogmaals dezelfde effectgrootte bij een voldoende power van 0,80. G*power berekend nu de benodigde steekproefgrootte.

Na de steekproefgroottes bepaald te hebben is via 'determine' in G*power te bepalen welke groepsgemiddelde de gewenste cohen's f oplevert. Er is gekozen om een standaarddeviatie van 25 te gebruiken voor alle groepen. Het streven was om deze standaarddeviatie ongeveer een kwart van het totale gemiddelde te laten zijn. De groepsgemiddelden zijn weergegeven in Tabel 1.

Met de groepsgemiddelde, standaarddeviatie, power, effectgrootte en steekproefgrootte bekend, is vervolgens met behulp van het programma 'BIEMS' bijbehorende exacte data gegenereerd (Mulder, Hoijsink & de Leeuw, 2012). Data Input < Select or Generate> <Generate data>. Bij Number of groups: 4. De rest blijft staan op default settings. (Number of dependent variables: 1, Number of explanatory variables: 0, ISeed: -1). <Use Settings>.

Nu worden de verschillende tabbladen doorlopen en ingevuld. Bij het tabblad <Group sizes> worden de steekproefgroottes in 4 gelijke delen ingevuld. Vervolgens het tabblad <Error covariance matrix specification> waar bij standard deviations 25 wordt ingevuld. De rest blijft op default settings (Correlation matrix: 1). Op het tabblad <Group means/regression coefficients> worden de gemiddelde van de groepen ingevuld zoals bepaald met behulp van G*power. Om exacte data te genereren wordt 'Generate exact data' aangevinkt en zal BIEMS een dataset genereren. Dit wordt herhaald voor de 6 condities zoals zijn weergegeven in tabel 1.

Met deze data zullen de volgende tests worden uitgevoerd in SPSS (IBM SPSS Statistics 24): ANOVA, Non-parametrische test (K&W), Bootstrap, contrast test met vaste contrast afstand, contrast test met relatieve contrasten en in het software programma 'R' zal met het package 'restriktor' (Vanbrabant, version 0.1-55) een F-bar test worden uitgevoerd. Daarnaast zullen deze datasets worden geanalyseerd met Bayesiaanse statistiek. Dit door gebruik te maken van de software programma's JASP (JASP Team, 2016). JASP (Version 0.7.5.5)[Computer software] Bij het programma JASP is bij <advanced options> de mogelijkheid om 'R scale fixed effect' te wijzigen naar respectievelijk 0,2 en 0,8 (default setting is 0,5) en bij BIEMS (Mulder, Hoijtink & de Leeuw, 2012) is het mogelijk om een gerichte hypothese te testen door een zelf een zogenoemd 'model' te construeren. In Appendix I zijn de scripts te vinden om de data te analyseren volgens de verschillende methodes.

Resultaten

In de analyses zoals weergegeven in tabel 2 is te zien dat de gegenereerde data met voldoende power van 0,8 bij veel van de gebruikte statistische methodes significante resultaten laten zien. Dit betekent dat bij voldoende power ook bij kleine datasets statistische significantie te bereiken is. Daarentegen zijn de resultaten verdeelt bij een lage power van 0,4. Opvallend is dat bij frequentistische statistiek de contrast testen en F-bar test die zijn uitgevoerd, wel sprake is van significante resultaten waar het voor de ANOVA, Non-parametrische en bootstrap test niet het geval is.

Bij Bayesiaanse statistiek is iets soortgelijks zichtbaar. Hier zijn in het software programma JASP, Bayes factoren >3 behaald bij voldoende power. Daarentegen zijn bij het programma BIEMS, waar een informatieve hypothese te testen is, voor alle datasets een Bayes factor >3 behaald. De resultaten zijn grafisch weergegeven in figuur 2 en 3.

In figuur 2 en 3 zijn de grenswaarden met een grijze lijn in het figuur aangebracht. Zo is inzichtelijk gemaakt wat de onderlinge verschillen in uitkomst zijn van de verschillende statistische methodes. Elke lijn representeert een andere dataset. De datasets met voldoende power (0,8) geven een redelijk consistent beeld waar de datasets met een lage power grotere verschillen laten zien.

Discussie

De statistische methode die gekozen wordt om data te analyseren is van belang. Deze statistische methode kan het verschil maken tussen het verwerpen of niet verwerpen van de nulhypothese. Door de gegenereerde data te analyseren met verschillende statistische methode is inzichtelijk gemaakt wat de gebruikte software in combinatie met gekozen statistische methode voor effect heeft op de resultaten en daarmee op de eindconclusie.

Zoals uit tabel 2 blijkt geven de ANOVA, Bootstrap en non-parametrische methode geen significante resultaten ($\alpha < 0,05$) bij een dataset met een lage power van 0,4. Bijzonder is wel dat de contrasttesten en F-bar test bij dezelfde data wel een significant resultaat opleveren. Dit verschil valt te verklaren omdat bij deze testen meer informatie is meegenomen doordat vooraf een gerichte uitspraak is gedaan over de resultaten in de vorm van een informatieve hypothese bij de F-bar test of gewichten zijn toegekend aan de groepen bij een contrast test. Door deze gerichte informatie mee te nemen in de gebruikte statistische methode is dus wel een significant onderzoeksresultaat behaald.

Wanneer gekeken wordt naar figuur 3 is te zien dat ook bij Bayesiaanse testen verschil te zien is in uitkomsten. Dit valt te verklaren doordat in BIEMS wel gewerkt kan worden met een informatieve hypothese en deze functie in JASP (nog niet) beschikbaar is. Je bent dus feitelijk andere hypothesen aan het testen. In JASP is het wel mogelijk om een betere of slechtere fit te geven in de vorm van de 'R scaled fixed effects'. Door een gunstige 'R scaled fixed effects' in te voeren kan verklaard worden waarom de datasets met een grote

power toch even boven een Bayes factor van 3 uitstijgen (figuur 3). In het programma BIEMS was het wel mogelijk een informatieve hypothese te testen, hierdoor is dus gerichte informatie toegevoegd en dit blijkt evenals bij de contrast testen en F-bar test te zorgen voor significante resultaten bij een lage power ($BF > 3$).

De aanbeveling van Van de Schoot, en Depaoli, (2014) is zichtbaar wanneer getest wordt met een informatieve hypothese. Zij vinden in hun artikel dat Bayesiaanse statistiek goed gebruikt kan worden bij kleine steekproeven omdat de statistische methode hier niet van afhangt. Het blijkt dat wanneer wordt getest met een informatieve hypothese alle datasets significante resultaten laten zien ($BF > 3$).

Het wel of niet significant zijn van een effect is arbitrair. Gebruikelijk in de sociale wetenschap is een alfa level van 0,05 te hanteren als cut-off punt bij frequentistische statistiek. Minder gebruikelijk is een cut-off punt voor een Bayes factor van 3 bij Bayesiaanse statistiek. Door Jefferys (1961) en Kass & Raftery (1995) is deze waarde van 3 gegeven als een positieve verbetering ten opzichte van de nulhypothese. Dit wordt vaak geciteerd om onderlinge vergelijkingen te kunnen maken tussen de frequentistische en Bayesiaanse statistiek. Naast power en effectgrootte is dus ook zeker het significantielevel van invloed op de conclusie .

Daarnaast hangt de eindconclusie ook zeker af van de keuze van statistische methode. Wat opvalt is dat wanneer er informatieve hypothesen worden getoetst, dat dit van invloed is op de p-waarden dan wel de Bayes factor. Welke keuze gemaakt wordt en hoe dan nu gerapporteerd gaat worden ligt nog open voor discussie.

Het is heel belangrijk om van tevoren te bedenken wat getest gaat worden, welke hypothesen hierbij horen en of er reeds informatie beschikbaar is. De statistische methode die gebruikt gaat worden zou afhankelijk moeten zijn van deze vragen. Er kan op basis van de

resultaten in deze thesis worden geadviseerd om echt na te denken over bestaande informatie en deze informatie vervolgens mee te nemen in de analyse.

Limitaties

In dit onderzoek is gekozen om gegenereerde data te gebruiken met een steekproefgrootte die overeenkomt met de gewenste waarde voor de parameters die zijn ingesteld. Dit maakt dat het misschien niet representatief is voor de onderzoekspraktijk toch ben ik van mening dat de resultaten wel de essentie van het werken met kleine datasets weergeven.

Een andere limitatie is het gebruik van cut-off waardes. Door deze in te stellen op een Bayes factor van 3 en een alfa level van 0,05 doen we eigenlijk geen recht aan de onderliggende theorie. In de praktijk zal altijd naar de individuele datasets gekeken moeten worden om het significantielevel te bepalen.

Aanbevelingen

Net als de navigatieofficier die zijn koers moet bepalen om in de juiste haven uit te komen, bepaalt de statisticus de methode die hij gaat gebruiken om zijn data te analyseren. Uit de resultaten blijkt dat wanneer van te voren wordt nagedacht over eventuele aanwezige informatie vooraf, dit de kans aanzienlijk vergroot om met een kleine steekproef toch statistische significantie te bereiken. Het verdient aanbeveling om voorafgaande aan het onderzoek dus na te denken of eventuele informatie aanwezig is en dit mee te nemen in de data-analyse.

Ten tijde van dit schrijven was het programma JASP voor Bayesian ANOVA nog in ontwikkeling. Hierdoor zijn in de toekomst misschien meer functies beschikbaar voor het vaststellen van een prior. Dit zal voor onderzoekers een gebruiksvriendelijke manier voor het analyseren van een Bayesiaanse ANOVA opleveren en daarmee makkelijker toegankelijk voor alle onderzoekers in het werkveld.

Voor vervolgonderzoek zou het mogelijk zijn om experts te ondervragen hoe zij in de praktijk deze data zouden rapporteren. Welke statistische methode kiezen de wetenschappers om te analyseren en te rapporteren en wat zijn hiervoor de argumenten.

Referentielijst

- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science*, 6(3), 274-290. Doi:10.1177/1745691611406920
- Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4e ed.). London, United Kindom: SAGE publications Ltd.
- Gravetter, F. J., & Wallnau, L. B. (2016). *Statistics for the behavioral sciences custom edition*. Hampshire, United Kingdom: Cengage Learning EMEA.
- Jeffreys, H. (1961). *Theory of probability* (3e ed.). Oxford, Groot-Brittannië: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773-795. Geraadpleegd van <http://www.jstor.org/stable/2291091>
- Kluytmans, A., Schoot, R. van de., Mulder, J., & Hoijtink, H. (2012). Illustrating Bayesian evaluation of informative hypotheses for regression models. *Frontiers in Psychology*, 3(2). doi:10.3389/fpsyg.2012.00002
- Kumar Dwivedi, A., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine*, 1-19. doi:10.1002/sim.7263
- Magnusson, K. (z.j.). Understanding Statistical Power and Significance Testing an interactive visualization. Geraadpleegd van <http://rpsychologist.com/d3/NHST/>
- Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, 46(2).
- Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative models for small samples in psychological research: applying linear mixed effects models and

generalized estimating equations to repeated measures data. *Educational and Psychological Measurement*, 76(1), 64-87. doi:10.1177/0013164415580432

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (ter perse).

Bayesian analysis of factorial designs. 1-57. Geraadpleegd van

<http://www.ejwagenmakers.com/inpress/RouderEtAlinpressANOVAPM.pdf>

Schoot, R. van de., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, 16(2), 75-84.

Schoot, R. van de., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A.G. van.

(2014). A gentle introduction to Bayesian analysis: Applications to developmental research.

Child development, 85(3), 842-860. doi:10.1111/cdev.12169

Vanbrabant, L., Schoot, R. van de, & Rosseel, Y. (ter perse). An introduction to restriktor:

informative hypothesis testing for AN(C)OVA and linear models. 1-49.

Vanbrabant, L., Schoot, R. van de, & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for ANOVA and regression. *Frontiers in Psychology*, 5(1565), 1-8.

doi:10.3389/fpsyg.2014.01565

Tabel 1

Gegenereerde exacte data in BIEMS

	Power 0,40	Power 0,80
Klein	n=456 (n=114 per groep)	n=1096 (n=274 per groep)
Cohen's $f = 0,10$	SD=25	SD=25
	M ₁ =112	M ₁ =112
	M ₂ =115	M ₂ =115
	M ₃ =117	M ₃ =117
	M ₄ = 119	M ₄ = 119
Middel	n=80 (n=20 per groep)	n=180 (=45 per groep)
Cohen's $f = 0,25$	SD=25	SD=25
	M ₁ =106	M ₁ =106
	M ₂ =107	M ₂ =107
	M ₃ =118	M ₃ =118
	M ₄ = 120	M ₄ = 120
Groot	n=36 (n=9 per groep)	n=76 (n=19 per groep)
Cohen's $f = 0,40$	SD=25	SD=25
	M ₁ =97	M ₁ =97
	M ₂ =108	M ₂ =108
	M ₃ =118	M ₃ =118
	M ₄ = 123	M ₄ = 123

Tabel 2

Analyse resultaten van de verschillende statistische methode

n=	n=36	n=76	n=80	n=180	n=456	n=1096
	Groot	Groot	Middel	Middel	Klein	Klein
Cohen's <i>f</i>	(0,4)	(0,4)	(0,25)	(0,25)	(0,10)	(0,10)
Power	0,4	0,8	0,4	0,8	0,4	0,8

Frequentistische statistiek

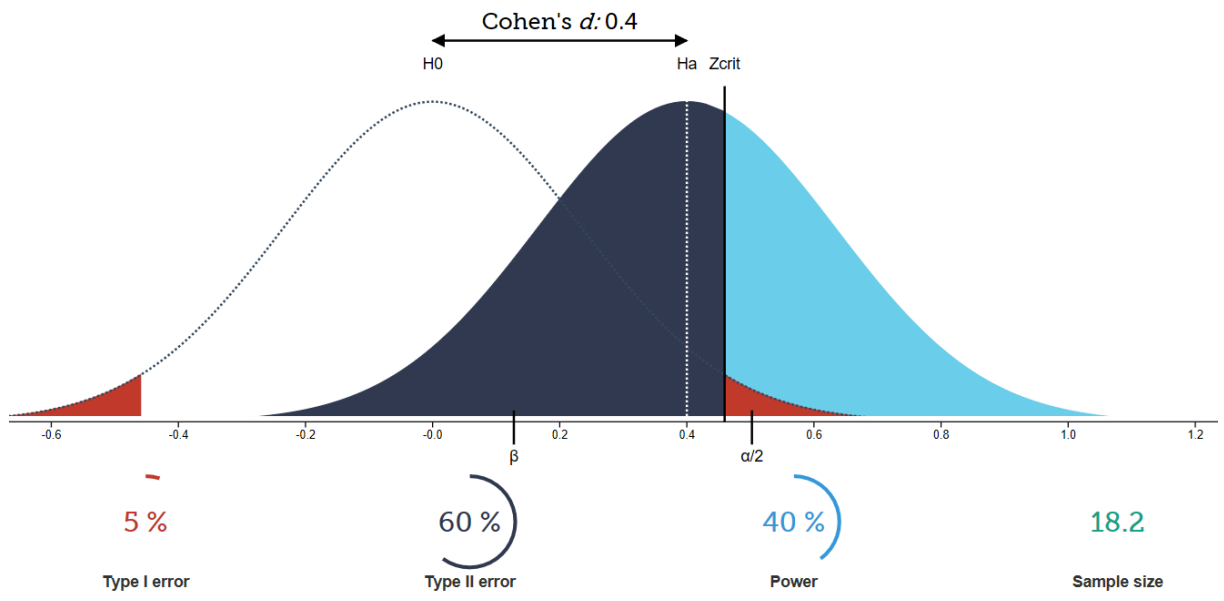
ANOVA	0,188	0,014	0,194	0,012	0,186	0,009
Non-Parametrisch (K&W)	0,183	0,024	0,155	0,051	0,179	0,017
Bootstrap	0,188	0,014	0,194	0,012	0,186	0,009
Contrast test (-1,5;- 0,5;0,5;1,5)	0,033	0,001	0,042	0,002	0,029	0,001
Contrast test (verhouding gem.)	0,031	0,001	0,031	0,001	0,028	0,001
F-bar Global	0,046	0,002	0,046	0,002	0,043	0,001
F-bar type C (t-distributie 1 side)	0,346	0,275	0,451	0,426	0,274	0,175

Bayesiaanse statistiek

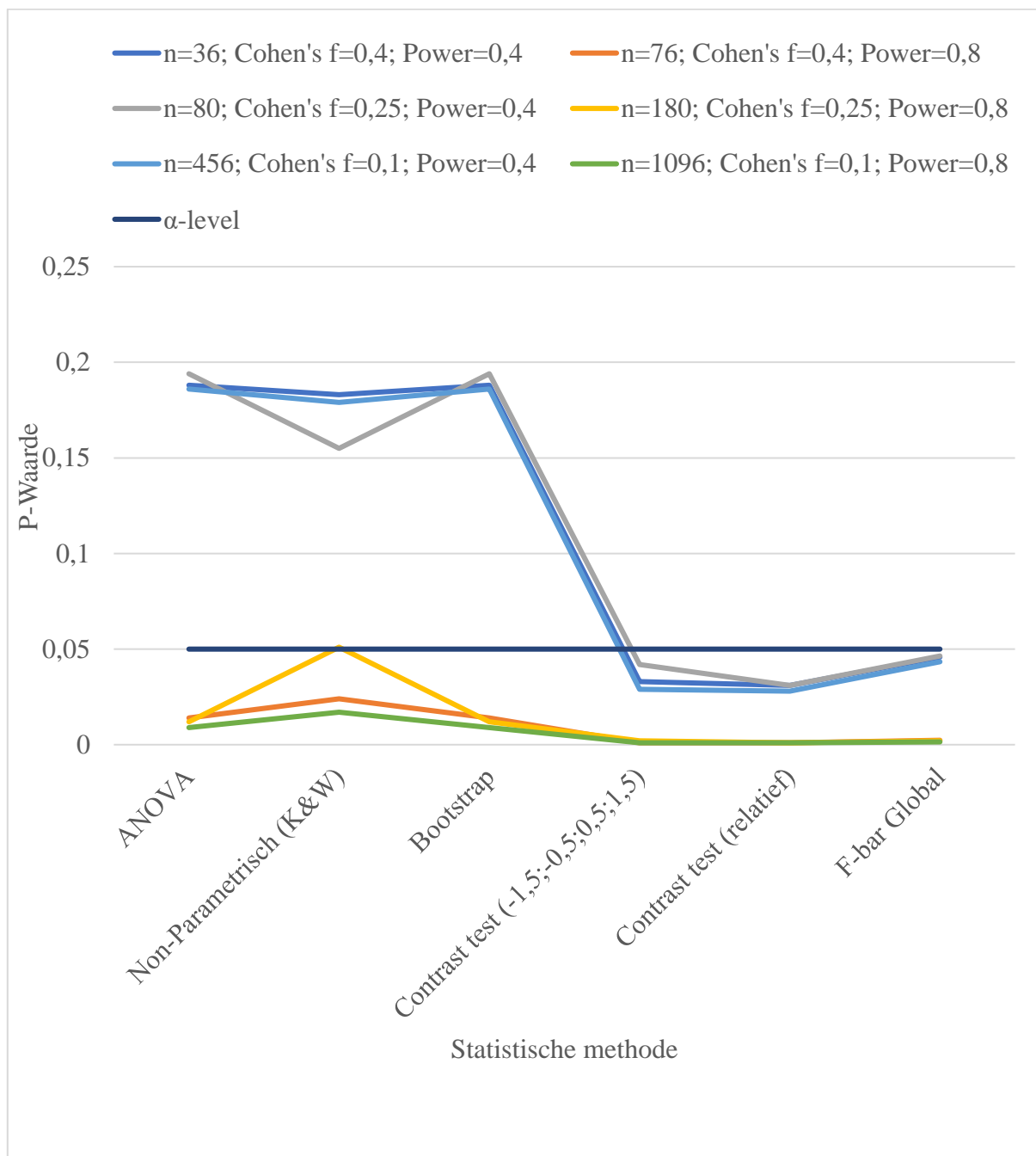
JASP (Default prior; r scale fixed effects = 0,5)	BF01:	BF01:	BF01:	BF01:	BF01:	BF01:
	1,671	0,276	2,731	0,394	13,102	1,620
	BF10:	BF10:	BF10:	BF10:	BF10:	BF10:
	0,598	3,621	0,366	2,535	0,076	0,617
JASP (set prior; r scale fixed effects = 0,2)	BF01:	BF01:	BF01:	BF01:	BF01:	BF01:
	1,089	0,253	1,289	0,232	2,567	0,292

	BF10:	BF10:	BF10:	BF10:	BF10:	BF10:
	0,918	3,956	0,776	4,302	0,390	3,426
JASP (set prior; r scale fixed effects = 0,8)	BF01:	BF01:	BF01:	BF01:	BF01:	BF01:
	2,744	0,420	5,697	0,814	42,103	5,016
	BF10:	BF10:	BF10:	BF10:	BF10:	BF10:
	0,364	2,383	0,176	1,228	0,024	0,199
BIEMS (default settings)	7,19	12,01	5,76	8,29	8,26	13,9
BF:10						

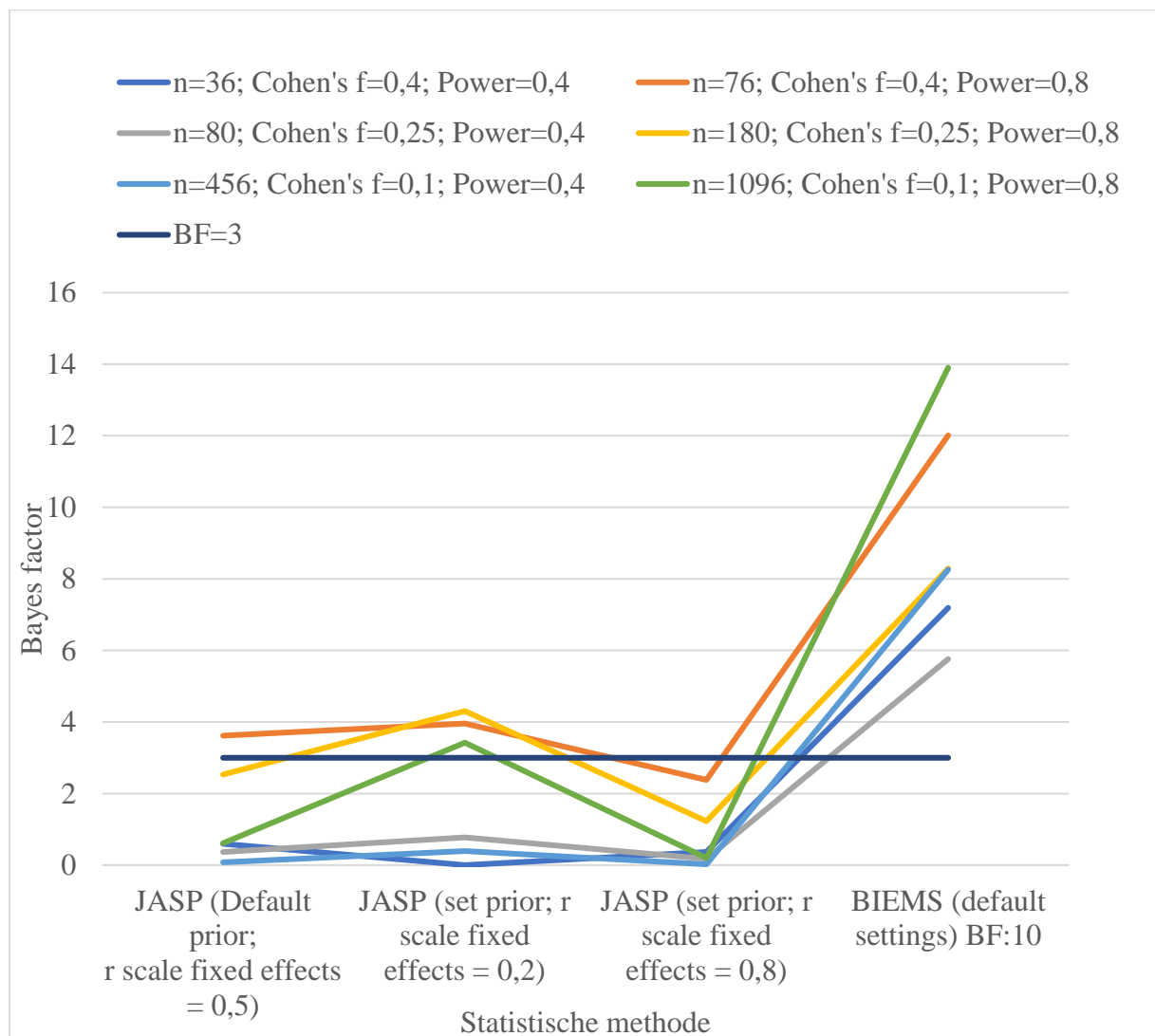
Figuur 1. Power grafisch weergegeven



Figuur 2. Frequentistische statistieken van de datasets bij verschillende statistische methodes



Figuur 3. Baysiaanse statistieken van de datasets bij verschillende statistische methodes



Appendix I

ANOVA syntax SPSS

De ANOVA analyse wordt uitgevoerd in SPSS.

```
DATASET ACTIVATE DataSet1.
```

```
ONEWAY Score BY Groep
```

```
/MISSING ANALYSIS.
```

Non-parametrisch (K&W) syntax SPSS

*Nonparametric Tests: Independent Samples.

```
NPTESTS
```

```
/INDEPENDENT TEST (Score) GROUP (Groep)
```

```
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
```

```
/CRITERIA ALPHA=0.05 CILEVEL=95.
```

Bootstrap syntax SPSS

```
BOOTSTRAP
```

```
/SAMPLING METHOD=SIMPLE
```

```
/VARIABLES TARGET=Score INPUT=Groep
```

```
/CRITERIA CILEVEL=95 CITYPE=PERCENTILE NSAMPLES=1000
```

```
/MISSING USERMISSING=EXCLUDE.
```

```
ONEWAY Score BY Groep
```

```
/MISSING ANALYSIS.
```

Contrast test (standaard contrast) syntax SPSS

ONEWAY Score BY Groep

```
/POLYNOMIAL=1
```

```
/CONTRAST=-1.5 -0.5 0.5 1.5
```

```
/MISSING ANALYSIS.
```

Contrast test (relatieve contrasten) syntax SPSS

Uit de gegenereerde data blijkt dat in dit voorbeeld dataset 'n=36' de volgende gemiddelden heeft: $M_1=97$, $M_2=108$, $M_3=118$, $M_4=123$. Voor deze contrast test zijn de contrasten op de volgende manier bepaald: $(4 * M_x) - (M_1 + M_2 + M_3 + M_4)$ Waarbij voor M_x respectievelijk de gemiddelde van 1, 2, 3 en 4 worden ingevuld om tot de relatieve contrasten te komen. Voor de dataset n=36 zijn de gebruikte relatieve contrasten: -58, -14, 26 en 46 en worden vervolgens geanalyseerd door SPSS:

ONEWAY Score BY Groep

```
/POLYNOMIAL=1
```

```
/CONTRAST=-58 -14 26 46
```

```
/MISSING ANALYSIS.
```

F-bar analyse

Hiervoor is gebruikgemaakt van het programma 'R'. Met het onderstaande script kan de data in 'R' worden geanalyseerd.

```
# package laden
```

```
library(foreign)
```

```
library(restriktor)
```

```
#file --> change dir --> goede map aanklikken (waar het bestand in zit, in dit voorbeeld heet  
het bestand 'n=36x.sav'.)
```

```
n36 <- read.spss("n=36x.sav", to.data.frame=TRUE)
```

```
# n36 is dan de naam van de dataset
```

```
View(n36) # voor het opvragen van de dataset, om te controleren of deze goed is ingevoerd.
```

```
n36$Groep<-factor(n36$Groep,levels=c("x1","x2","x3","x4"))
```

```
myConstraints1<-'Groep1 < Groep2; Groep2 < Groep3; Groep3 < Groep4'
```

```
fit_ANOVA <-lm(Score~-1+Groep,data=n36)
```

```
ihf(fit_ANOVA, constraints = myConstraints1)
```

Bayesiaanse statistiek

JASP

```
#Laden van de dataset in JASP
```

```
#Tabblad Common
```

```
#Knop: ANOVA → selecteer Bayesian ANOVA
```

```
#Dependent variable = Score; Fixed factor = groep
```

```
# advanced options → 'R scale fixed effect' te wijzigen naar respectievelijk 0,2 en 0,8 (default  
setting is 0,5
```

```
#Door te selecteren valt te wisselen tussen BF10 en BF01
```

BIEMS

#<Select or Generate>, <existing data> laadt hier de dataset

#Vul bij number of dependent variables 1 in. <Ok>

specificeer model door $\mu(1,1) < \mu(2,1)$, $\mu(2,1) < \mu(3,1)$ en $\mu(3,1) < \mu(4,1)$ in te voeren.

<define as model>

#Dit wordt nu “Model 1” genoemd en is je informatieve hypothese.

selecteer bij <standardize> Variable: Yes

#<Edit Model>

#Ga naar stap 3 Generate Default Prior <OK>

#Ga naar stap 4 <Calculate Bayes factors>

#<Calculate> Nu wordt de Bayes factor gegeven van de informatieve hypothese tegen de nul hypothese.