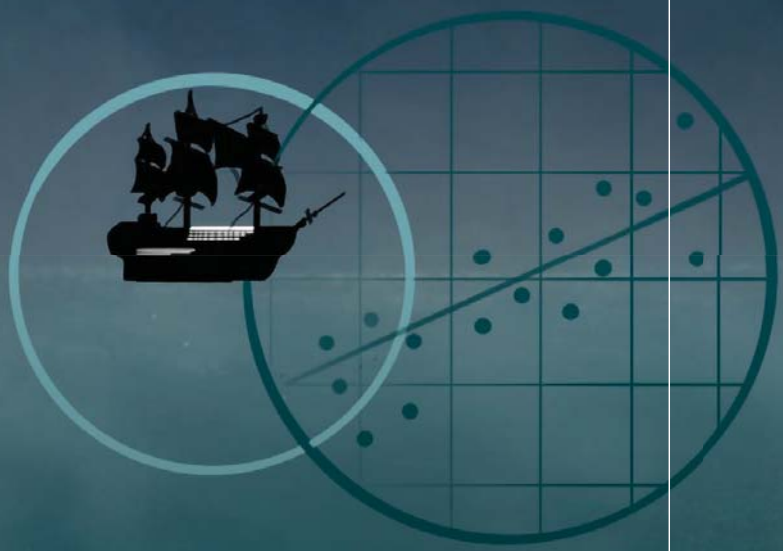
Statistiek in Stijl 2

Piraten, Perziken en P-waarden



Noordhoff

Vince Penders

1e druk

Statistiek in stijl 2

Vince Penders

Eerste druk

Noordhoff Groningen/Utrecht

Ontwerp omslag: Shootmedia, Groningen
Omslagillustratie: Chelsy Penders

Eventuele op- en aanmerkingen over deze of andere uitgaven kunt u richten aan: Noordhoff Uitgevers bv, Afdeling Hoger Onderwijs, Antwoordnummer 13, 9700 VB Groningen of via het contactformulier op www.mijnnoordhoff.nl.

0 / 21

ISBN (ebook) 978-90-01-29315-4
ISBN 978-90-01-29314-7
NUR 916

Zij die gemaakt hebben, groeten u

Welkom, beste lezer! Voor je neus ligt een splinternieuw handboek, en wat zijn we trots op het resultaat. Als makers willen wij ons kort aan je voorstellen. Mag dat?

Die jongeman links is **Vince Penders** … noem me maar **Vincenzo**. Het brein achter deze lesmethode – dat ben ik. Creatief tot op het bot, een vrolijke perfectionist. Vroeger was ik werkzaam bij de Universiteit Maastricht, maar tegenwoordig werk ik als onafhankelijk docent en geef jaarlijks les aan grofweg 300 studenten. In de tussentijd verbeter ik constant mijn lesboeken. Op de zeldzame momenten waarop ik niet werk of onder vrienden ben, sta ik te koken, een Nintendospel te spelen of fictie te schrijven met epische muziek in mijn oren. Jawel: verhalen zijn mijn tweede (of eigenlijk eerste) passie. Mijn meest recente roman, *De beer in het brein*, vertelt het relaas van een man die geloofde dat hij iets kleins en pluizigs was. Evenals mijn debuut *Zwaluwhart* is het uitgebracht door Macc.

Mijn zus **Chelsy Penders** (vooraan) zingt, danst en werkt als interieurstylist. Waar mijn artistieke stem het geschreven woord is, is de hare de handgemaakte tekening in elke vorm en stijl. Chelsy's veelzijdigheid blijkt wel uit de vele illustraties die dit boek rijk is, evenals uit de gelikte kleurenschema's (vergeef me mijn gebrek aan trots op haar werk). Haar Maine

Coon-katten worden graag knuffelig als ze thuis zit te tekenen, dus laat het ons weten als je ergens een kattenpoot tegenkomt.

En dit is **Sophia von Stockert**, een oud-student van me en tegenwoordig een goede vriendin. Als toegewijd statistiekdocent inspireerde ze vroeger studenten om te ontdekken wat voor coole dingen je allemaal met statistiek en onderzoek kunt doen. Om deze reden heb ik haar aan boord gehaald als redacteur van *Statistiek in stijl*. Sophia heeft een diploma in de Research Master Cognitive and Clinical Neuroscience en volgt momenteel een opleiding tot psychotherapeut in Heidelberg. In haar vrije tijd klimt ze graag en ze redigeert en vertaalt boeken en manuscripten.

Dan nog een paar eervolle vermeldingen als dankbetuigingen. De eerste is voor Olav Wessel, die het lef had om mijn manuscript telefonisch bij Noordhoff aan te bieden en ons vervolgens wist binnen te loodsen via een obscuur online formulier van de klantenservice. De tweede gaat naar Vincent Diks, mijn uitgever, die vanaf het eerste moment in dit boek geloofde. Ten slotte een dikke omhelzing voor alle studenten die me elke dag inspireren. Jullie helpen is een eer. Jullie hebben me ertoe overgehaald dit boek te schrijven en mijn passie brandende gehouden om het verder te verbeteren met deze nieuwe editie. Mensen bedanken me vaak voor het werk dat ik doe, maar er bestaat slechts één juist antwoord. ‘Nee … *jullie* bedankt!’

Inhoud

Hoe gebruik je dit boek? 7
Overzichten en kernbegrippen 9

DEEL 1
Data beschrijven 33

1 Dataverkenning voor gevorderden 35

DEEL 2
Variantieanalyse 79

2 Vergelijkingen en contrasten 81

3 Tweeweg-ANOVA 121

4 ANCOVA 171

5 RANOVA 203

6 Tweeweg-RANOVA 249

7 Split-plot-ANOVA 273

DEEL 3
Categorieën en kruistabellen 307

8 Kruistabelanalyse 309

DEEL 4
Regressie 353

9 Dichotome en dummyvariabelen 355

10 Meervoudige regressie 389

10A Hoofdeffecten 391

10B Interactie en simpele hellingen 437

10C Mediatie 463

11 Logistische regressie 477

11A Hoofdeffecten 479

11B Interactie-effecten 503

DEEL 5
Psychometrie 525

12 Factoranalyse 527

13 Betrouwbaarheid 581

14 Validiteit en diagnostiek 629

DEEL 6
Bruggen slaan 655

15 Contrasten en ANOVA 657

16 ANOVA en regressie 661

17 Kruistabelanalyse en logistische regressie 677

Appendix 682

Begrippenlijst Nederlands-Engels 697

Begrippenlijst Engels-Nederlands 698

Register 700

Hoe gebruik je dit boek?

Statistiek in stijl kan fungeren als een op zichzelf staande cursus statistiek. In de praktijk zullen de meeste lezers het waarschijnlijk gebruiken ter ondersteuning van een vak op de hogeschool of universiteit. Daarom hebben we het boek opgesplitst in *twee delen*:

- *Statistiek in stijl 1* dekt de basis af. Steekproeven beschrijven, kanstheorie, hypothesetoetsen, betrouwbaarheidsintervallen … Waarschijnlijk heb je iets aan dit deel als je een *hogeschoolstudent* bent, of in je *eerste jaar van de universiteit* zit.
- *Statistiek in stijl 2* opent het volledige arsenaal. Statistische modellen met meer dan twee variabelen, herhaalde metingen, psychometrie … Zet koers naar deze wateren in je *tweede jaar van de universiteit en verder*.

Uiteraard is dit slechts een richtlijn. Individuele curricula bespreken mogelijk wat onderwerpen uit boek 2 op de hogeschool of komen terug op stof van boek 1 gedurende het tweede universiteitsjaar. Onderzoekers (en perfectionistische studenten) kunnen het als prettig ervaren om beide boeken te bezitten, zodat ze de kleine details kunnen nalezen. Merk op dat een bezitter van enkel boek 2 zich geen zorgen hoeft te maken: allerlei sleutelbegrippen uit boek 1 staan samengevat in de paragraaf *Overzichten en kernbegrippen*! Gebruik deze voor een soepele studeerervaring. ☺
Elk hoofdstuk kent zijn eigen moeilijkheidsgraad en vereist soms voorkennis uit eerdere hoofdstukken. Dit staat telkens aangegeven op de introductiepagina. De moeilijkheidsgraden zijn als volgt:

Niveau	Toelichting
 Cupcake	Voor de nieuwkomer in de wereld van de statistiek. Geen wiskundeknobbel? Geen probleem! Ook jij gaat na een dag studeren met een gerust hart naar bed.
 Zeemeeuw	Hier wordt het wat spannender. Geduld en herhaling zijn de ingrediënten waarmee je de eindstreep haalt. Moeilijker dan dit gaan we het in boek 1 niet maken.
 Gorilla	Het echte werk. Veel wetenschappelijk onderzoek is op dit soort statistiek gebaseerd. De meeste hoofdstukken in boek 2 zijn van gorilla-niveau.
 Ninja	Expert in de dop, hier moet je wezen. Neem je je vak serieus, dan leren deze hoofdstukken je behoorlijk geavanceerde technieken. Voorkennis en motivatie zijn een must. In ruil daarvoor wordt statistiek cooler dan ooit tevoren.

Behalve theoretische uitleg bevat elk hoofdstuk ook *opdrachten*. Ik heb deze zo ontworpen dat ze je kennis niet slechts op de proef stellen, maar ook verdiepen. Oefenen dus! Goede opdrachten zijn leerzaam en voegen veel toe aan je praktische ervaring. In de meeste hoofdstukken vind je maar liefst twee setjes. De *opdrachten voor perziken* zijn het fijnst om mee te beginnen; de *opdrachten voor piraten* zijn wat uitdagender.
O, en de uitwerkingen? Die kun je gratis downloaden van de website, *statistiekinstijl.noordhoff.nl*.

Ben je zelfstandig bezig met statistiek?
Begin dan gewoon bij hoofdstuk 1 of het onderwerp waarover je graag iets wilt leren. Zie het volgende kopje voor bijzondere afwegingen.

Volg je een statistiekvak op de hogeschool of universiteit?
Ga naar de inhoudsopgave en zoek het onderwerp op dat je moet leren. Kun je het niet vinden, raadpleeg dan eens de index achter in het boek. Onderwerpen met een *groene markering* bespreek ik ook in de *Overzichten en kernbegrippen*-paragraaf.
Gevonden? De voorkennis die je nodig hebt, staat telkens aan het begin van het hoofdstuk vermeld. Lees eerst met een goede kop koffie of thee de theorie en maak aantekeningen. Schrijf een eigen samenvatting, vergelijk deze op het eind met die van mij en verbeter je eigen exemplaar waar nodig. Maak daarna de opdrachten om je nieuwgeboren kennis op de proef te stellen. De uitwerkingen vind je gratis op de website.
Raak vooral niet ontmoedigd als de opdrachten je niet meteen goed afgaan. Statistiek heeft aandacht en oefening nodig. Soms gaat het een dag later al veel beter, als de nieuwe stof een beetje is bezonken. Veel succes! ☺

Heb je statistiek nodig voor je eigen onderzoek?
Waarschijnlijk ben jij al bekend met de onderwerpen die ik in dit boek bespreek. Je bent echter het een en ander vergeten of wilt wat details controleren. In dat geval is het van belang dat je gemakkelijk je weg kunt vinden. Bestudeer in elk geval het *Stappenplan voor analysen* in de appendix – het vertelt je niet alleen wat je moet doen, maar ook *hoe je de resultaten in APA-stijl kunt rapporteren*. Zodra je de statistische analyse hebt gekozen die bij je onderzoek past, kun je verder naar het bijbehorende hoofdstuk. In de online leeromgeving vind je tevens instructies om SPSS aan te sturen. Mocht je hiernaast je kennis van een specifiek begrip willen opfrissen, dan biedt de paragraaf *Overzichten en kernbegrippen* je ondersteuning. Ik hoop dat je begeleider onder de indruk zal zijn van jouw statistische arsenaal.

Overzichten en kernbegrippen

De stof die in dit boek aan bod komt, kent veel centrale concepten. Deze komen telkens terug. Dit overzicht bespreekt ze een voor een, en wil dan ook een basis zijn waarop je altijd kunt terugvallen. Gebruik het volgende niet om nieuwe dingen te leren, maar keer er steeds naar terug als je iets bekends vergeten bent. Wanneer in de hoofdstukken een begrip *groen* is gemarkeerd, verwijst het je naar deze proloog.

Assumpties *Geïntroduceerd In boek 1, hoofdstuk 7*

Alle statistische modellen en hypothesetoetsen maken een aantal aannames. Dat zijn dingen waar het model of de toets van uitgaat bij de berekeningen, zodanig dat de uitkomsten alleen ergens op slaan als die assumpties correct zijn.

¿Que? zullen vele lezers nu denken, dus laat ik het concreter maken. Neem een standaard keukenweegschaal. Met dit instrument kan ik een zak zoete aardappelen wegen (of paprika's, als die meer jouw ding zijn). Het apparaat zal me vertellen dat mijn zoete aardappelen 1,2 kilo wegen. Maar de keukenweegschaal gaat ervan uit dat ik hem op een bepaalde plaats gebruik … Nee, niet in de keuken. Wijsneus. Ik bedoel de planeet Aarde. Ik zou naar de maan kunnen reizen en mijn weegschaal daar gebruiken; ongetwijfeld verschijnt er ook dan een aantal kilogrammen op het scherm. Dit getal zal echter compleet *onjuist* zijn, omdat de weegschaal afgesteld is op het zwaartekrachtniveau van onze planeet. Het punt is dat mijn instrument me altijd een uitkomst geeft, maar slaat die uitkomst ergens op? Alleen als er aan de assumpties van het instrument is voldaan.

Met statistische toetsen gaat het net zo. Zij leunen op bepaalde assumpties, en we kunnen de toetsresultaten alleen vertrouwen onder die voorwaarden. Echter! Er bestaan uitzonderingen die kunnen voorkomen dat de toetsen uit elkaar vallen. Uitzonderingen die de toetsen *robuust* maken, zoals statistici zeggen. Let in de hoofdstukken op de kleurcodes: een rood gedrukte assumptie kan niet worden gepasseerd, een blauw gedrukte assumptie wel.

Robuuste analyse

Betrouwbaarheidsinterval *Geïntroduceerd in boek 1, hoofdstuk 6*

Om een beeld te krijgen van een populatieparameter, zoals een gemiddelde, doen we er goed aan een betrouwbaarheidsinterval op te stellen. Steekproefschattingen zijn namelijk nooit perfect nauwkeurig. In hoofdstuk 6 van *Statistiek in stijl 1* mat men bijvoorbeeld een steekproef van aliens van de planeet Coronyyyr om te bepalen hoeveel armen de gemiddelde Coronyyyriaan had. Het gemiddelde aantal armen in de steekproef was 4,27. Het populatiegemiddelde kon echter iets afwijken. Daarom maakten we een 95%-betrouwbaarheidsinterval: we plaatsten een foutmarge rond het steekproefgemiddelde, om te bepalen waartussen het populatiegemiddelde zou moeten liggen. Het interval was [3,11; 5,47].

De berekening van betrouwbaarheidsintervallen kent een vaste structuur, ongeacht om wat voor een parameter het gaat:

steekproefschatter ± *kritieke waarde* × *standaardfout*

De kritieke waarde is een z- of t-waarde, die aangeeft *hoeveel* standaardfouten we hanteren voor de foutmarge. (Voor een korte herhaling van de standaardfout, zie *Steekproevenverdeling*.) Een hogere kritieke waarde resulteert in een hoger betrouwbaarheidsniveau (zeg, niet 95% maar 99%): het interval wordt breder, dus we kunnen met meer zekerheid stellen dat het de populatieparameter binnen zijn grenzen weet te vatten. Een breder interval is echter ook minder precies. Om het weer te versmallen zonder de betrouwbaarheid op te offeren, kun je proberen een grotere steekproef te trekken – hetgeen de standaardfout verkleint.
Betrouwbaarheidsintervallen laten zich ook gebruiken als tweezijdige *hypothesetoetsen*. Kijk nog eens naar het interval voor de aliens op Coronyyyr. Als het populatiegemiddelde inderdaad tussen 3,11 en 5,47 ligt, kunnen we iets zeggen over een keur aan nulhypothesen. Zo is 2 armen *geen* aannemelijk populatiegemiddelde; de nulhypothese dat $\mu = 2$ moet worden verworpen. H_0: $\mu = 3$ zouden we eveneens verwerpen. 4 of 5 daarentegen zijn aannemelijke waarden voor het populatiegemiddelde, dus we kunnen de nulhypothesen niet verwerpen dat $\mu = 4$ of $\mu = 5$. In essentie vertelt het betrouwbaarheidsinterval je over *alle nulhypothesen die je niet zou verwerpen* (bij een significantieniveau van 1 min het gekozen betrouwbaarheidsniveau: $\alpha = 1 - C$).

Bonferroni-correctie *Geïntroduceerd in boek 1, hoofdstuk 10*

Soms bekijken we verbanden tussen variabelen met meerdere toetsen achter elkaar. Bijvoorbeeld in de volgende situaties:

- *Paarsgewijze vergelijkingen:* het vergelijken van meer dan twee groepen op dezelfde afhankelijke variabele. In boek 1, hoofdstuk 10 bestudeerden we drie behandelingen om meerminnen meer tevreden te maken over hun lichaam. Uiteindelijk vergeleken we alle condities een-op-een, met drie aparte t-toetsen.
- *Geplande contrasten:* in hoofdstuk 2 bestudeerden we vier behandelingen die vampiers moesten helpen meer slachtoffers te maken. We contrasteerden diverse clusters van groepen, en twee van deze contrasten waren *niet orthogonaal* (zie elders).
- *Simpele-effectenanalysen:* in hoofdstuk 3 ontdekten we dat geheim agenten slechter poker speelden als ze het opnamen tegen een crimineel (vergeleken met een potje tegen eerlijke burgers). Echter, dit effect van de tegenstander was minder sterk aanwezig bij agenten die zelfverzonnen cocktails mochten drinken – vermoedelijk omdat zij sowieso al slechter presteerden. Dus na het vinden van dit significante *interactie-effect* (zie verderop) ligt het voor de hand dat we het tegenstandereffect bekijken voor beide cocktailgroepen apart: de groep die eigen cocktails mocht bedenken en de groep die de drankkaart van het casino moest aanhouden. We doen dat (in dit geval) met twee eenweg-ANOVA's of twee t-toetsen.

Aan deze aanpak kleeft echter een probleem. Elke keer dat je een statistische toets uitvoert, loop je het risico een *type I-fout* te maken (zie elders). Door meerdere toetsen te verrichten, maak je de kans op *minstens één*

type I-fout dus groter dan normaal. Draai je drie toetsen, dan kan die kans soms uitgroeien tot bijna 15% (aangenomen dat $\alpha = 0{,}05$ voor elke toets). We noemen deze kans de *family-wise error rate*: de mate waarin er een fout kan worden gemaakt in een 'familie' van toetsen.
De Bonferroni-correctie is een basale methode om de *family-wise error rate* terug te brengen naar 5%. Ze bestaat eruit dat we het significantieniveau delen door het aantal toetsen:

$$p \quad vs. \quad \frac{\alpha}{\#toetsen}$$

Het significantieniveau wordt dus verlaagd. P-waarden moeten nu worden vergeleken met het aangepaste significantieniveau. Is een p-waarde nog steeds lager dan dit nieuwe significantieniveau, dan verklaren we de toets significant. Dit maakt het moeilijker om de nulhypothese te verwerpen, en minder waarschijnlijk dat we hem verwerpen als-ie *waar* is (type I-fout).
Je kunt de Bonferroni-correctie ook toepassen door het significantieniveau onaangeroerd te laten, en in plaats hiervan elke p-waarde te vermenigvuldigen met het aantal toetsen:

$$p \times \#toetsen \quad vs. \quad \alpha$$

Effectief is dit hetzelfde als de eerste methode (de linker- en rechterkant van de vergelijking zijn gewoon vermenigvuldigd met het aantal toetsen). Het levert daarom precies dezelfde resultaten op. Ik ben gesteld geraakt op deze tweede aanpak, maar het is een kwestie van smaak.
Een paar opmerkingen tot slot:

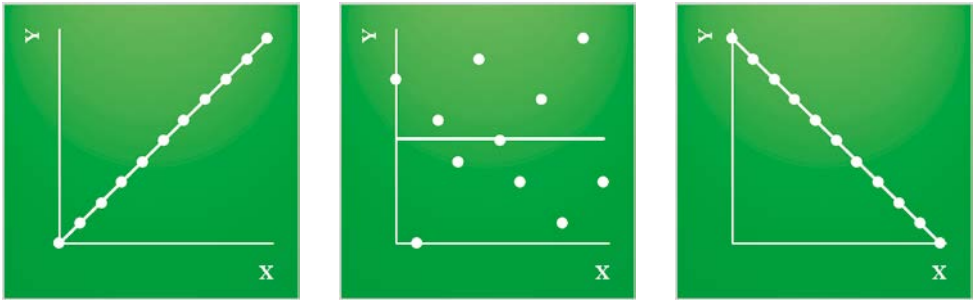
- Voor paarsgewijze vergelijkingen bestaan er nog vele andere correctiemethoden. Om eerlijk te zijn ben ik geen enorme fan van het leeuwendeel (ze zijn abstract en de meeste geven je niet veel power in ruil voor de moeite). Check hoofdstuk 2, subparagraaf 2.1.III, voor wat suggesties.
- Voor gevorderden is de *aangepaste Bonferroni-correctie* – naar mijn mening – een fraaie variatie op de basale Bonferroni. Ik raad hem mijn hogerejaarsstudenten altijd aan. Zie hoofdstuk 2, subparagraaf 2.1.II.

Centrale-limietstelling	Centreren	Cohens *d*
Zie *Steekproevenverdeling*	Zie *Datatransformatie*	Zie *Effectgrootte*

Correlatie *Geïntroduceerd in boek 1, hoofdstuk 3*

Als we het lineaire verband willen beschrijven tussen twee kwantitatieve variabelen, kunnen we gebruikmaken van een correlatiecoëfficiënt. Dit is een gestandaardiseerde maat; hij neemt altijd een waarde aan tussen –1 en 1. Figuur 0.1 helpt je bij de interpretatie. Als r_{XY} ongeveer 0 is, is het verband tussen de twee variabelen erg zwak of zelfs afwezig. Hoe verder r_{XY} afwijkt van 0, des te sterker is het verband. Hij kan zoals gezegd minimaal –1 zijn; in dat geval spreken we van *perfecte negatieve samenhang*. Als de correlatiecoëfficiënt 1 is (maximaal), is er *perfecte positieve samenhang*.

FIGUUR 0.1 Positieve correlatie ($r = 1$), geen correlatie ($r = 0$) en negatieve correlatie ($r = -1$)



Je zou de lineairecorrelatiecoëfficiënt zelf kunnen berekenen, maar dan kom je er wel achter waarom we tegenwoordig een computer het werk laten doen. 😐 De meest gebruikelijke formule is:

$$r_{XY} = \frac{covariantie_{XY}}{s_X s_Y} = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})/(N-1)}{s_X s_Y}$$

Deze geldt voor steekproeven. Voor een uitwerking kun je kijken in hoofdstuk 3 van *Statistiek in stijl 1*.

Data-indeling *Geïntroduceerd in boek 1, hoofdstuk 8 (SPSS-instructies)*
Bezig met je eigen onderzoek? Zorg ervoor dat je het databestand correct opzet. Als je dit vanaf het begin doet, scheelt het een hoop pijn in de bips. Datasets laten zich op een paar manieren indelen en dat bepaalt het soort analyses waaraan je de data kunt onderwerpen in programma's zoals SPSS. Alle SPSS-instructies in de online leeromgeving vermelden de juiste indeling voor de bijbehorende analyse.

Univariate (lange) indeling

TABEL 0.1 Univariaat databestand (voorbeeld)

Persoon	Beloning	Redding (*Y*)
1	1	97
1	2	91
1	3	94
2	1	74
... (etc.)	...	...

Een voorbeeld van een univariate dataset ziet eruit zoals in tabel 0.1. De eerste rij bevat de reddingsscore van proefpersoon 1 in beloningsconditie 1, de tweede rij bevat de reddingsscore van dezelfde persoon in beloningsconditie 2, enzovoort. Iedere deelnemer bezet dus meerdere rijen en we hebben *één kolom* voor *één enkele afhankelijke variabele*. Alle basale *between-subjects-designs* werken met deze indeling, omdat iedere deelnemer slechts één keer gemeten wordt.

Multivariate (brede) indeling
Multivariate databestanden worden met name gebruikt voor *herhaalde metingen* oftewel *within-subjects-designs*. Ze benadrukken dat elke deelnemer meermaals gemeten is, doordat ze eruitzien alsof die deelnemer een score heeft op meerdere afhankelijke variabelen:

TABEL 0.2 Multivariaat databestand (voorbeeld)

Persoon	Redding: medaille (Y_1)	Redding: geld (Y_2)	Redding: film (Y_3)
1	94	91	97
2	42	49	74
... (etc.)	...	...	...

Hier heeft iedere proefpersoon één rij waarin we drie scores vinden, één per 'afhankelijke' variabele (dat zijn er drie, in verschillende kolommen). Merk op: de onafhankelijke variabele van het onderzoek, beloning, is min of meer opgegaan in de diverse 'afhankelijke' variabelen Y.
Natuurlijk zijn multivariate datasets ook een vereiste voor analysen die daadwerkelijk *meerdere afhankelijke variabelen* omvatten, zoals een MANOVA (niet besproken in dit boek).

Gewogen gevallen

TABEL 0.3 Databestand met gewogen gevallen (voorbeeld)

Toekomst	Aantal
1	90
2	18
3	12

Onderzoeken met enkel *categorische variabelen* kunnen werken met een verkorte dataset. Je moet SPSS vertellen elke rij te wegen op basis van de aantal-variabele (onder *[Data] [Weight Cases]*). Vanaf dan ziet het programma niet slechts één persoon die toekomst 1 gescoord heeft (getrouwd), maar 90 personen. Tevens zijn er 18 personen met toekomst 2 (gescheiden) en 12 met toekomst 3 (dood).
Merk op dat je dezelfde gevens ook in een univariate indeling kunt verwerken, maar dan moet je 90 rijen maken met een 1 erin, 18 rijen met een 2 erin en ga zo maar door.

Datatransformatie *Geïntroduceerd in boek 1, hoofdstuk 1*
Wie bijvoorbeeld de lengte van een aantal bananen heeft gemeten in meters, kan ervoor kiezen om alle scores om te zetten in centimeters (dat lijkt mij ook wat handiger in dit geval). Hij of zij voert dan een datatransformatie uit: alle gegevens worden omgezet naar andere waarden, maar hun betekenis blijft hetzelfde. De lengte van de bananen verandert tenslotte niet als je van 0,2 meter naar 20 centimer gaat.
Het simpelste soort datatransformaties zijn *lineaire transformaties*: multipliceren, centreren en standaardiseren (naar *z-scores*; zie verderop). Ik

behandel ze in hoofdstuk 1 van *Statistiek in stijl 1*. Daarnaast is er een gecompliceerdere verzameling genaamd *monotone transformaties*, zoals worteltrekken, de logaritme nemen en inverteren. Die staan besproken in hoofdstuk 1 van dit boek.

Effectgrootte *Geïntroduceerd in boek 1, hoofdstuk 6*

Als een effect significant is (wat erop duidt dat het werkelijk bestaat in de populatie), vertelt dit niet het hele verhaal. Is het effect best groot of heel klein? Deze vraag is minstens even belangrijk. Zo werd in hoofdstuk 7 van *Statistiek in stijl 1* de nulhypothese verworpen: het bleek dat vrouwen die Pavlov-conditioner gebruiken hun haar inderdaad langer dan 48 uur in een aantrekkelijke staat kunnen houden … maar hoelang? Het bleek rond de 49 uur te liggen, hetgeen nauwelijks meer dan 48 is.
Er zijn heel wat grootheden ontwikkeld om effectgrootten uit te drukken. Vele hiervan zijn helaas zo abstract dat ze in de echte wereld weinig betekenis hebben. Ik houd dingen graag simpel. In tabel 0.4 vind je een aantal relatief laagdrempelige indicatoren van effectgrootten. Hoofdstuk 1 van dit boek gaat verder in op een aantal geschikte *schatters*.
Opmerking: de richtlijnen voor de effectgrootte zijn zeer algemeen. Jacob Cohen, de bedenker, gaf al aan dat specifiekere richtlijnen per wetenschapsveld kunnen verschillen. Bijvoorbeeld: in de psychologie zou een correlatie van 0,20 al gelden als medium en 0,30 als groot. Nog grotere effecten zijn in dit domein namelijk erg zeldzaam.

TABEL 0.4 Overzicht van effectmaten en schatters

Maat	Gebruik	Richtlijnen	Geschikte schatter(s)
Delta (δ) Gestandaardiseerd verschil tussen gemiddelden (z-score)	T-toetsen (om 2 groepen per keer te vergelijken)	♦ 0,20: klein ♦ 0,50: medium ♦ 0,80: groot	*Cohens d* (uitgaand van gelijke varianties), Hedges' g (idem) of *Glass'* $\hat{\Delta}$ (niet uitgaand van gelijke varianties).
Èta-kwadraat (η^2) Proportie verklaarde variantie	ANOVA	♦ 0,01: klein ♦ 0,06: medium ♦ 0,14: groot	$\hat{\eta}^2$ is vrij onzuiver en overschat de proportie verklaarde variantie systematisch. Gebruik bij voorkeur $\hat{\omega}^2$ of (nog beter) $\hat{\varepsilon}^2$.
Odds-ratio (ζ) Verhouding tussen twee odds	Kruistabellen, logistische regressie	♦ 1,68 (0,60): klein ♦ 3,47 (0,29): medium ♦ 6,71 (0,15): groot	De steekproef-OR is een *zuivere schatter* (zie verderop) van ζ.
Correlatie (ρ) Sterkte van lineair verband	Lineaire regressie	♦ 0,10: klein ♦ 0,30: medium ♦ 0,50: groot	De steekproefcorrelatie r is een tikje onzuiver, maar wordt verreweg het meest gebruikt. Beter is de wortel uit de *aangepaste* R^2.

Maat	Gebruik	Richtlijnen	Geschikte schatter(s)
R-kwadraat (ρ^2) Proportie verklaarde variantie	Lineaire regressie	♦ 0,01: klein ♦ 0,09: medium ♦ 0,25: groot	De steekproef-R^2 is vrij onzuiver en overschat de proportie verklaarde variantie systematisch. Gebruik bij voorkeur de *aangepaste* R^2.

Efficiënte schatter	Family-wise error rate	Hoofdeffect
Zie *Schatter*	Zie *Bonferroni-correctie*	Zie *Interactie, hoofdeffecten en simpele effecten*

Hypothesetoetsing *Geïntroduceerd in boek 1, hoofdstuk 6*

Heel wat sociologische, medische en economische onderzoeken werken met steekproeven. Zo kreeg in hoofdstuk 10 van *Statistiek in stijl 1* een steekproef van meerminnen de vraag voorgelegd hoe tevreden ze zich voelden met hun eigen lichaam, nadat een willekeurig samengestelde groep een chirurgische vinvergroting had ondergaan, een tweede willekeurig samengestelde groep therapie had gekregen om een immaterialistische levenshouding te ontwikkelen, en een derde groep twee maanden op een wachtlijst had gestaan. Het gemiddelde lichaamsbeeld bleek tussen de drie groepen te verschillen … maar dit kan ook komen door de samenstelling van de steekproef. Trek een nieuwe en de resultaten zullen waarschijnlijk iets veranderen. Hypothesetoetsen (en betrouwbaarheidsintervallen; zie een eerdere paragraaf in *Overzichten en kernbegrippen*) proberen boven de onzekerheid van steekproeven uit te stijgen, en te bewijzen dat de steekproef ons ook iets vertelt over de populatie.
Vier stappen vormen de basis van elke significantietoets:

I Stel de hypothesen op

De *nulhypothese* van elke toets beweert dat het effect dat ons interesseert *niet bestaat in de populatie*. In de meermincasus hield dit in dat de experimentele behandeling niet in staat was het lichaamsbeeld te veranderen:

$$H_0: \mu_{vinvergroting} = \mu_{groepstherapie} = \mu_{controle}$$

De *alternatieve hypothese* stelt dat het effect op een bepaalde manier wel bestaat:

$$H_A: niet\ alle\ \mu_i\ gelijk$$

Als we erin slagen de nulhypothese te verwerpen, tonen we aan dat de alternatieve waarschijnlijk waar is in zijn plaats. Dit geeft ons statistisch bewijs voor het effect dat we zoeken.

II Trek de steekproef

Bestudeer wat voor effect(en) je vindt op steekproefniveau. Hoogstwaarschijnlijk zul je iets tegenkomen. Zoals gezegd liet de meermincasus in *Statistiek in stijl 1* enige verschillen zien: de vinvergrotingsgroep had gemiddeld het hoogste tevredenheidsniveau (jammer genoeg), gevolgd door

de therapiegroep, en de controlegroep kwam als laatste. Dat is niet raar, want elke conditie bevatte 29 meerminnen. Het zou juist verrassend zijn geweest als die drie vrij kleine groepen zich exact hetzelfde gedroegen. Echter, elk verschil dat je vindt kan om deze reden *toeval* zijn. Het is nog steeds mogelijk dat de experimentele behandeling totaal geen effect had, en dat de groepen simpelweg verschilden doordat ze verschillende individuen bevatten. Was dit het geval? In een poging om het uit te sluiten …

III Bereken de p-waarde

De p-waarde is de *kans dat je een steekproefeffect zou vinden zoals je gevonden hebt – of een nog groter effect – indien de nulhypothese waar is.*
Een hoge p-waarde duidt erop dat enkel steekproeftoeval een rol heeft gespeeld. Stel dat de p-waarde gelijk is aan 0,43. Als de nulhypothese in dat geval waar is (de experimentele behandeling had dus geen effect), zou je zo'n verschil als jij tussen de groepen hebt gevonden, nog steeds 43% van de keren vinden dat je steekproeven trekt van deze omvang. Dit geeft ons geen krachtig bewijs dat de behandeling effectief is geweest: het steekproef-'effect' is waarschijnlijk maar toevallig, want het manifesteert zich vrij vaak wanneer je een aselecte steekproef trekt, zelfs als de nulhypothese juist is.
Een lage p-waarde daarentegen – die in feite uit dit onderzoek naar voren kwam: $p = 0{,}00007$ – wekt argwaan jegens de nulhypothese. Hij duidt aan dat, als de behandeling geen effect had, de kans slechts 0,007% is dat we in de steekproef een effect zoals dit zouden observeren. Het lijkt dan aannemelijker dat de onderzoeker helemaal niet zo'n zeldzame steekproef heeft getrokken, en dat de nulhypothese simpelweg niet klopt.
Is de p-waarde die je krijgt dus laag genoeg?

IV Beslis

In feite moet je een *significantieniveau* (α) kiezen voorafgaand aan stap II. Dit is de grens voor de p-waarde: als die uitkomt gelijk aan het significantieniveau of lager, zul je de nulhypothese verwerpen – anders niet. Het is verstandig om een significantieniveau aan te houden dat gebruikelijk is in jouw wetenschapsveld. In *Statistiek in stijl* zal ik standaard 0,05 oftewel 5% hanteren (behoudens Bonferroni-correcties et cetera).

TABEL 0.5 De beslisregel bij hypothesetoetsen

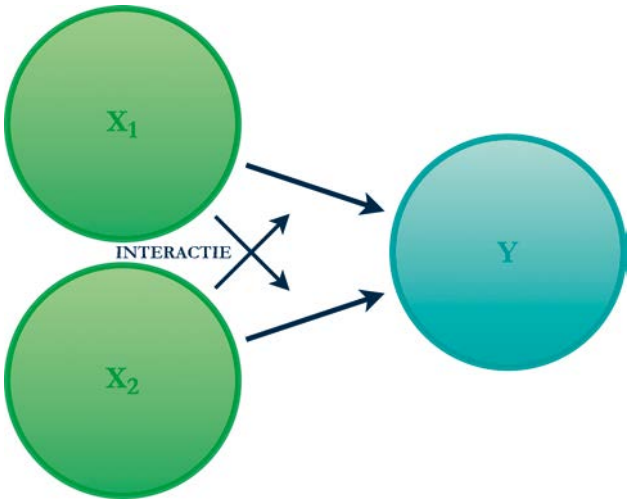
$p > \alpha$	Als de nulhypothese klopt, is dit niet zo'n bijzondere steekproef. Het resultaat is *niet significant*: het geeft niet aan (signaleert niet) dat de nulhypothese niet juist is. Deze kan daarom *niet* worden *verworpen*.
$p \leq \alpha$	Als de nulhypothese klopt, is dit een vrij bizarre steekproef. Het resultaat is *significant*: het geeft aan (signaleert) dat de nulhypothese niet juist is. Deze kan daarom worden *verworpen*.

De conclusie van het meerminnenonderzoek was dat de nulhypothese verworpen kon worden, en dat de behandeling dus een echt effect had in de meerminnenpopulatie. (Tip: zie ook het artikel over type I- en type II-fouten. Een fout was hier niet waarschijnlijk, aangezien de p-waarde extreem klein was.)

Interactie, hoofdeffecten en simpele effecten *Geïntroduceerd in hoofdstuk 3, 8 en 10*

Met statistische modellen zoals meerweg-ANOVA en meervoudige regressie onderzoeken we de effecten van *meer dan één* onafhankelijke variabele. In hoofdstuk 3 vroegen we ons af of geheim agenten die de vrijheid krijgen om zelfverzonnen cocktails te bestellen, beter pokeren tegen criminelen. Cocktails (X_1) en tegenstander (X_2) waren de onafhankelijke variabelen in dit onderzoek, en de hoeveelheid fiches die de agenten wonnen (Y) was de afhankelijke. Het is mogelijk dat elke onafhankelijke variabele op zichzelf de fiches beïnvloedt. Maar het is ook mogelijk dat ze elkaars effecten beïnvloeden – dus verzwakken of versterken. In het eerste geval hebben we alleen *hoofdeffecten*, maar in het tweede geval spreken we van *interactie*.

FIGUUR 0.2 Interactie

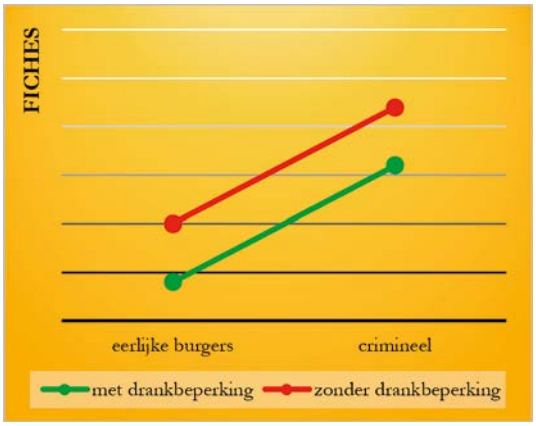


Het effect van X_1 op Y hangt af van X_2; omgekeerd hangt het effect van X_2 op Y dan automatisch af van X_1 (figuur 0.2). Daarom kunnen we tweeweginteractie uitbeelden met een kruis – ze werkt altijd twee kanten op. (Ik zeg altijd graag dat interactie symmetrisch is.) Maar hoe moeten we ons de aan- of afwezigheid van dit effect concreet voorstellen? Hierna vind je een paar voorbeeldgrafieken. De gewonnen fiches staan op de y-as en de punten zijn gemiddelden.

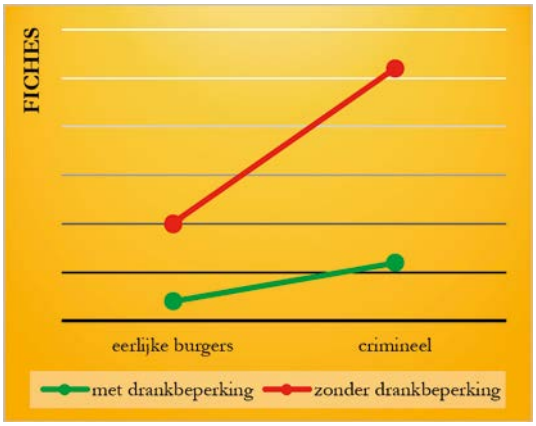
In figuur 0.3 spelen agenten gemiddeld altijd beter tegen een crimineel, of ze nu qua cocktailconsumptie beperkt worden of niet. Wie weet verhoogt het belang van het toernooi hun concentratie. Daarnaast ligt de rode lijn van de groep zonder beperking consequent boven de groene lijn van de groep met beperking. Dit is een hoofdeffect van de factor cocktails: agenten met meer vrijheid presteren beter, mogelijk omdat ze zich meer op hun gemak voelen. Hun tegenstander verandert daar niets aan.

Figuur 0.4 laat iets anders zien: daar is het verschil tussen de tegenstandergroepen klein als de agenten beperkt worden in hun drankconsumptie, maar groot als ze mogen bestellen wat ze willen. Anders gezegd, het effect van de tegenstander op het aantal gewonnen fiches (de afhankelijke variabele Y) verandert onder invloed van cocktails. Dit noemen we een *interactie-effect*.

FIGUUR 0.3 Hoofdeffecten



FIGUUR 0.4 Interactie



Interactie doet zich dus voor wanneer we het tegenstandereffect niet in het algemeen bestuderen, maar *apart* voor agenten met en zonder drankbeperking, en we vervolgens merken dat er een verschil is. Elke lijn in figuur 0.3 en 0.4 toont een *simpel effect* van de tegenstander. In het algemeen definiëren we simpele effecten als de effecten van een factor B, *per niveau van factor A*.
(Een hoofdeffect daarentegen is een *gewogen gemiddelde* van de simpele effecten (de gewichten zijn gebaseerd op steekproefgrootten). In figuur 0.3 en 0.4 kunnen we het hoofdeffect van de tegenstander zien als het midden tussen de twee lijnen.)

Controleren op interactie

- In ANOVA-modellen kun je het best een grafiek (laten) maken zoals figuur 0.3 en 0.4, en kijken of de lijnen *parallel* lopen. Zo ja, dan is er *geen interactie*. Ook al zal zich in steekproeven waarschijnlijk minstens een beetje interactie voordoen, je moet altijd eerst nagaan of die significant is alvorens je een conclusie trekt.
  Pas op: de grafiek moet gebaseerd zijn op groeps*gemiddelden* op de afhankelijke variabele. Verwar dit niet met een frequentietabel, die het

aantal proefpersonen aangeeft in elke combinatiegroep. Frequenties beschrijven slechts de verdeling van de *onafhankelijke* variabelen, en zijn onder meer relevant om *orthogonaliteit* na te gaan.

- In regressiemodellen is checken op interactie nog makkelijker: neem een interactieterm op in het model en zie of zijn regressiecoëfficiënt gelijk is aan 0 of niet. Wederom moet je tevens rekening houden met zijn p-waarde en/of betrouwbaarheidsinterval.

De volgende stap

- Heb je geen significante interactie gevonden? Analyseer *hoofdeffecten*. Het effect dat de tegenstander heeft op de pokerprestaties is altijd even sterk, ongeacht iemands cocktailconsumptie. Wil je hoofdeffecten bekijken, dan kan het nodig zijn om eerst de interactieterm weg te halen uit je statistische model.
  (Dit laatste is telkens nodig als je onafhankelijke variabelen niet orthogonaal zijn (zie elders). Je moet het altijd doen in regressiemodellen, tenzij je predictoren niet alleen orthogonaal zijn maar ook gecentreerd (zie hoofdstuk 10B). Bij twijfel, gewoon verwijderen! ☺)
- Heb je een significante interactie gevonden? Analyseer *simpele effecten*. Het effect dat de tegenstander heeft op de pokerprestaties verandert door een drankbeperking, dus het is niet zo productief om het overkoepelende hoofdeffect te bestuderen. Interactie kan er zelfs voor zorgen dat hoofdeffecten hun betekenis verliezen. In figuur 0.5 is het hoofdeffect positief: in het algemeen winnen de agenten meer fiches tegen een crimineel. Maar kijken we naar de losse simpele effecten, dan brokkelt de betekenis van het hoofdeffect af. Hoe bedoel je, er is een algemeen effect van de tegenstander? Niet onder agenten met een drankbeperking!

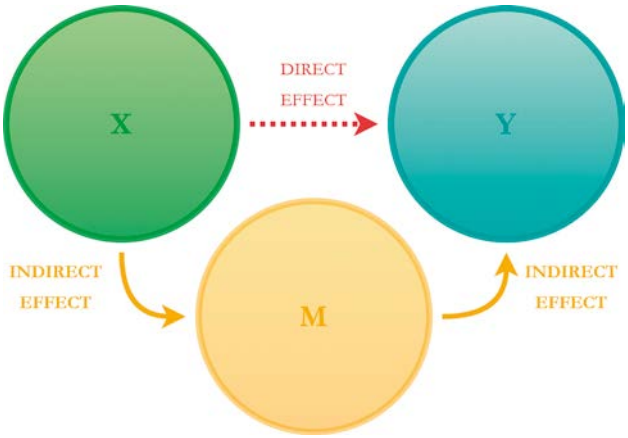
FIGUUR 0.5 Interactie, maar geen zinvol hoofdeffect van de tegenstander

Kwadratensom	Mean square
Zie *Variantie*	Zie *Variantie*

Mediatie *Geïntroduceerd in hoofdstuk 10C*

Dit statistische effect onderscheidt zich van interactie en verstoring. In hoofdstuk 10C ontdekten we dat de hoeveelheid rum die piratenbendes dronken een negatief effect had op hun welvaart, gemeten in dukaten. Waarom zouden drinkende piraten doorgaans minder dukaten bezitten? Een verklaring kan zijn dat drinken een piratenbende ongeorganiseerd maakt; dit slechte bestuur kon hen vervolgens geld kosten (bijvoorbeeld doordat ze fouten maakten in hun administratie). De variabele *BESTUUR* bevindt zich dus in het midden van het causale pad tussen *RUM* en *DUKATEN*, en hij *verklaart* hoe dit causale pad tot stand komt (figuur 0.6). We noemen hem daarom een *mediator*.

FIGUUR 0.6 Mediatie



Er zijn diverse methoden ontwikkeld om mogelijke mediatie te inspecteren. Zie hoofdstuk 10C.

Non-parametrische toetsen *Niet elders besproken*

Soms zijn gangbare statistische toetsen zoals ANOVA of regressie misschien geen goed idee. Je kunt ze niet zonder meer gebruiken *als de steekproeven die je getrokken hebt erg klein zijn* en je niet durft te vertrouwen op een of meerdere *assumpties* (vaak normaliteit). Om dit soort situaties het hoofd te bieden, zijn een paar alternatieve toetsen ontwikkeld. De term 'non-parametrisch' is een ietwat verkeerde benaming; hoewel ze niet kijken naar gemiddelden ($\overline{Y}$), kijken ze wel naar *medianen*. Hier volgt een beknopt overzicht van de vier meest wijdverspreide non-parametrische (schmon-parametrische) technieken:

I Wilcoxon-rangtekentoets

Deze functioneert als een alternatief voor de *gepaarde t-toets* (zie boek 1, hoofdstuk 8). We gebruiken hem dus voor een *within-subjects*-design (herhaalde metingen) of gekoppelde paren. Net zoals de gepaarde t-toets kijkt de rangtekentoets naar de verschilscores V van alle proefpersonen, en hij toetst de nulhypothese dat hun *mediaan gelijk is aan nul* in de populatie:

$$H_0: M_{V,populatie} = 0$$

In plaats van $\mu_V = 0$, het populatiegemiddelde. Als de mediaan hoger of lager ligt dan 0, hebben de twee condities die we vergelijken de neiging om te verschillen op de afhankelijke variabele.
Vervolgens geven we de verschilscores in de steekproef een *rangnummer* van de laagste tot de hoogste, waarbij we negeren of ze positief of negatief zijn (de *V*'s dicht bij 0 krijgen dus een lage rang en de *V*'s ver weg van 0 krijgen een hoge). Verschilscores gelijk aan 0 gooien we weg. Daarna kijken we alleen naar de positieve *V*'s, en we *tellen hun rangen op* om een zogenoemde *rangsom* te berekenen. We doen hetzelfde met de negatieve *V*'s. De gedachte is dat als $M_V = 0$, dan zouden hoge en lage rangen gelijkmatig verdeeld moeten zijn over de positieve en negatieve sets van verschilscores; derhalve moeten de positieve en negatieve rangsom gelijk zijn. Hoe groter het verschil, des te meer dit duidt op een mediaan die afwijkt van 0. Als een van de rangsommen *significant* groter wordt dan de andere, dan mogen we de nulhypothese verwerpen.
De Wilcoxon-rangtekentoets gaat na of dit het geval is. Willen we hem gebruiken, dan moet er aan drie assumpties zijn voldaan:

- De afhankelijke variabele is op zijn minst ordinaal (anders kunnen we de verschilscores niet rangschikken).
- De groepen zijn afhankelijk (deze toets werkt alleen voor een within-subjects-design).
- De steekproef is getrokken uit een *symmetrisch* verdeelde populatie. Dit heeft iets weg van de normaliteitsassumptie, maar het is wat minder strikt; zie hoofdstuk 1 over het begrip kurtosis. Toch levert het nogal een paradox op: was de normaliteitsassumptie, die op de symmetrie-assumptie lijkt, niet juist een voorname reden waarom we onze toevlucht tot een non-parametrische toets hebben genomen? Bij een kleine steekproef laat die assumptie zich moeilijk controleren, maar als de steekproef groot is, mogen we ook gewoon een gepaarde t-toets gebruiken. Het behoort tot de redenen waarom ik persoonlijk niet zo warmloop voor deze non-parametrische alternatieven. Maar goed, bij kleine steekproeven en invloedrijke uitschieters kan de rangtekentoets toch meer valide zijn dan de t-toets.

Overigens is de rangtekentoets ook in staat de *t-toets voor 1 steekproef* te vervangen (zie boek 1, hoofdstuk 7). De truc is om als referentiepunt niet 0 te gebruiken, maar de gehypothetiseerde populatiemediaan (zeg 48, als we het voorbeeld uit hoofdstuk 7 volgen). Je moet dan rangen toekennen op basis van hoe ver scores afwijken van dit referentiepunt.

II Tekentoets
Een versimpelde vorm van de rangtekentoets. Deze versie negeert de grootte van de verschilscores: ze bekijkt alleen of een verschilscore positief is of negatief. Hiermee verliest de toets vaak een hoop power, maar hij gooit ook de assumptie van een symmetrische verdeling het raam uit. Gebruik bij twijfel dus de tekentoets. Hij toetst dezelfde nulhypothese, rust op dezelfde assumpties minus symmetrie, en werkt grotendeels op dezelfde manier.

III Wilcoxon-rangsomtoets / Mann-Whitney-U-toets
De twee namen verwijzen naar exact dezelfde toets. Deze is bedoeld om de *ongepaarde t-toets* te vervangen (zie boek 1, hoofdstuk 9): we gebruiken

hem voor *between-subjects*-designs. Hij heeft een vergelijkbare nulhypothese:

$$H_0: M_{1,populatie} = M_{2,populatie}$$

Dus dat beide populaties dezelfde *mediaan* hebben, in plaats van hetzelfde gemiddelde. Wederom behelst de toets dat we alle scores een rang geven, en dan de som van de rangen berekenen voor elke groep. Wederom is de gedachte dat die rangsommen ongeveer hetzelfde zouden moeten zijn; als dat het geval is, zijn hoge en lage rangen vrij gelijkmatig verdeeld over beide groepen, en dan hebben ze ongeveer dezelfde mediaan. Hoe groter het verschil tussen de rangsommen, des te waarschijnlijker is het dat de ene populatie een hogere mediaan heeft dan de andere.
De Wilcoxon-rangsomtoets neemt drie dingen aan:

- De afhankelijke variabele is op zijn minst ordinaal (anders kunnen we de scores niet rangschikken).
- De groepen zijn onafhankelijk (deze toets werkt alleen voor een between-subjects-design).
- De twee steekproeven zijn getrokken uit populatieverdelingen met *dezelfde vorm*. Opnieuw heeft dit wat weg van de normaliteitsassumptie. De paradox is dat deze assumptie bij kleine steekproeven lastig na te gaan is, maar dat we bij grote steekproeven net zo goed een ongepaarde t-toets kunnen gebruiken.

IV Kruskal-Wallistoets
Een uitbreiding van de rangsomtoets voor als we *twee of meer groepen* hebben. Hierdoor staat de Kruskal-Wallistoets naast de *eenweg-ANOVA* (zie boek 1, hoofdstuk 10). Hij toetst dezelfde nulhypothese als de rangsomtoets (mogelijk betreffende meer medianen), is gebaseerd op de identieke maat van rangsommen, en kent dezelfde assumpties. Enkel de toetsingsgrootheid is anders (het is een chi-kwadraatwaarde). Ik kan me zo voorstellen dat wanneer de Kruskal-Wallistoets significant blijkt, je hem kunt opvolgen met *paarsgewijze vergelijkingen* in de vorm van rangsomtoetsen, om te zien welke groepen precies verschillen van elkaar. Vergeet dan geen Bonferroni-correctie!

Algemene nadelen
Persoonlijke mening: ik heb het niet zo op non-parametrische toetsen. Daar kan ik de volgende redenen voor geven:

- Rangnummers maken de toetsen heel abstract. Je raakt gemakkelijk de draad kwijt.
- De meeste toetsen maken nog steeds een aanname over de vorm van de populatieverdeling. Het probleem dat ze behoren op te lossen, blijft daardoor slechts half opgelost.
- De toetsen zijn alleen geschikt voor heel basale designs. Zodra je een onderzoek uitvoert met meerdere onafhankelijke variabelen (en geloof me, dat gebeurt om de haverklap), kun je ze niet meer gebruiken.

Vandaar dat ik besloten heb geen heel hoofdstuk over deze technieken te schrijven; het universum van de statistiek is schier oneindig en dit boek heeft andere prioriteiten. Hopelijk helpt dit stukje je toch op weg, mocht je ooit meer willen weten over dit non-parametrische spul.

Orthogonaliteit *Geïntroduceerd in hoofdstuk 3, 8 en 10*

In een onderzoek met meer dan één onafhankelijke variabele moet je altijd in acht nemen of die onafhankelijke variabelen *orthogonaal* zijn. Simpel gezegd betekent dit dat ze *geen verband* vertonen. Orthogonale variabelen zijn prettig om te hebben, want die kunnen elkaar niet *verstoren* (zie de paragraaf over verstoring verderop). Hoe je hun verband nagaat, hangt af van hun meetniveaus:

TABEL 0.6 Zo check je op orthogonaliteit

X_1	X_2	Methode	Hoe waarschijnlijk is orthogonaliteit?
Categorisch (factor)	Categorisch (factor)	Maak een *kruistabel*. Zie hierna voor een verdere uitleg.*	Het is mogelijk om factoren compleet orthogonaal te maken als je een experiment uitvoert. Onwaarschijnlijk in observationeel onderzoek.
Categorisch (factor)	Kwantitatief (covariaat)	Check of de groepen van de factor hetzelfde *gemiddelde* hebben op de covariaat.	Onwaarschijnlijk in observationeel onderzoek. In een experiment zullen de factorgroepen niet *exact* hetzelfde gemiddelde hebben op de covariaat, maar als het goed is wel ongeveer (dankzij de willekeurige toewijzing van deelnemers). Dus de x-variabelen zullen niet compleet orthogonaal zijn, maar het gaat geen pijn doen.
Kwantitatief (covariaat)	Kwantitatief (covariaat)	Bereken de *lineaire correlatie* (die moet 0 zijn). Deze methode werkt ook als minstens één van de variabelen dichotoom is (dus 2 niveaus heeft).	Onwaarschijnlijk (kwantitatieve predictoren zijn typisch voor observationeel onderzoek). Hoe sterker de correlatie, hoe minder orthogonaal je x-variabelen.

* Pas op: dit moet een *frequentietabel* zijn – een die het aantal personen weergeeft in elke combinatiegroep. Verwar dit niet met een tabel van *gemiddelden*, die je de gemiddelde *scores op de afhankelijke variabele* geeft in elke combinatiegroep (als je afhankelijke variabele kwantitatief is).

In hoofdstuk 8 onderzocht Granger bijvoorbeeld of volbloed magiërs (van wie beide ouders konden toveren) vaker loyaal waren aan de Zwarte Zijde. Dit kon betekenen dat ze een inherente neiging hadden om toe te geven aan de verlokkingen van het kwaad. Echter, het was ook mogelijk dat volbloed tovenaars gewoon conservatiever waren, en dat dit een betere verklaring vormde voor hun keuze om zich aan te sluiten bij het kwaad. In de volgende voorbeelden zouden bloedstatus en conservatisme orthogonaal ofwel *gebalanceerd* zijn geweest:

TABEL 0.7A Orthogonaal design (voorbeeld 1)

		Bloedstatus (X_1)		
		gemengd bloed	volbloed	
Conservatisme (X_2)	progressief	50 (50%)	50 (50%)	**100 (50%)**
	conservatief	50 (50%)	50 (50%)	**100 (50%)**
		100 (100%)	**100 (100%)**	**200 (100%)**

TABEL 0.7B Orthogonaal design (voorbeeld 2)

		Bloedstatus (X_1)		
		gemengd bloed	volbloed	
Conservatisme (X_2)	progressief	40 (40%)	40 (40%)	**80 (40%)**
	conservatief	60 (60%)	60 (60%)	**120 (60%)**
		100 (100%)	**100 (100%)**	**200 (100%)**

TABEL 0.7C Orthogonaal design (voorbeeld 3)

		Bloedstatus (X_1)		
		gemengd bloed	volbloed	
Conservatisme (X_2)	progressief	60 (40%)	20 (40%)	**80 (40%)**
	conservatief	90 (60%)	30 (60%)	**120 (60%)**
		150 (100%)	**50 (100%)**	**200 (100%)**

Het gaat hier om de *relatieve-frequentieverdelingen* in de tabel: die behoren aan elkaar gelijk te zijn. In beide bloedstatusgroepen is de verhouding tussen progressieven en conservatieven hetzelfde (50-50 in tabel 0.7a en 40-60 in tabel 0.7b en 0.7c). Derhalve zijn de variabelen *statistisch onafhankelijk*: een willekeurig getrokken volbloed maakt net zoveel kans om conservatief te zijn als een willekeurig getrokken halfbloed.
(Dit betekent vanzelf dat beide politieke groepen dezelfde verhouding kennen van halfbloed en volbloed magiërs. Met andere woorden, het werkt ook andersom. Bereken de rijpercentages en je ziet het!)
Als de relatieve-frequentieverdelingen niet gelijk zijn, is het design *non-orthogonaal* oftewel *ongebalanceerd*. Neem tabel 0.8. Hierin waren de volbloed magiërs veel vaker conservatief dan degenen met gemengd bloed. De kans op het trekken van een conservatief verandert afhankelijk van de genetische groep; in een non-orthogonaal design hebben we geen statistische onafhankelijkheid. Zien we in zo'n geval dat volbloed tovenaars vaker de overstap maakten naar de Zwarte Zijde, dan wordt het moeilijker om te zeggen of dit een aangeboren karaktertrek van hen was of dat hun sociaal conservatisme de oorzaak vormde. Anders gezegd: conservatisme wordt mogelijk een verstoorder.

TABEL 0.8 Non-orthogonaal design

		Bloedstatus (X_1)		
		gemengd bloed	volbloed	
Conservatisme (X_2)	progressief	134 (72,0%)	56 (53,3%)	**190 (65,3%)**
	conservatief	52 (28,0%)	49 (46,7%)	**101 (34,7%)**
		186 (100%)	**105 (100%)**	**291 (100%)**

Pas op: sommige statistici noemen een design alleen orthogonaal/gebalanceerd wanneer alle celfrequenties *exact* hetzelfde zijn (tabel 0.7a). Ik gebruik de term een beetje losser. Dat doe ik omdat verstoring in al 'mijn' orthogonale situaties onmogelijk is.

Parameter *Geïntroduceerd in boek 1, hoofdstuk 5*

Uitspraak: 'parámeter'! ☺ Een vaste grootheid, zoals een gemiddelde, een standaardafwijking of een correlatie, die doorgaans aan een populatie toebehoort. Het zijn altijd de parameters die we echt willen weten. Jammer genoeg is het vaak niet mogelijk om een hele populatie te meten. In dat geval proberen we van parameters een indruk te krijgen met behulp van *steekproefschatters* (zie de *Schatters*-paragraaf). Ter aanduiding van parameters kunnen we het best *Griekse* symbolen gebruiken.

Power	P-waarde
Zie *Type I-fout, type II-fout en power*	Zie *Hypothesetoetsing*

Schatter *Geïntroduceerd in boek 1, hoofdstuk 5*

Een grootheid, zoals een gemiddelde, een standaardafwijking of een correlatie, berekend in een steekproef. We gebruiken deze steekproefschatter om de bijbehorende *populatieparameter* te schatten. Een paar voorbeelden:

- Met een steekproefgemiddelde, $\overline{X}$, schatten we het populatiegemiddelde μ.
- Met een steekproefcorrelatie, r_{XY}, schatten we de correlatie ρ_{XY} op populatieniveau.
- Met een regressiecoëfficiënt in de steekproef, b, schatten we de populatiecoëfficiënt β.

Zoals je ziet duiden we een schatter altijd aan met een *Latijns* symbool. Aangezien de waarde van een schatter steeds kan veranderen wanneer we een nieuwe steekproef trekken, is elke schatter een *toevalsvariabele* met zijn eigen *steekproevenverdeling*.

- Een schatter is *zuiver* indien zijn verwachtingswaarde gelijk is aan de parameter. Hij zou dus gemiddeld de parameter moeten opleveren als we de steekproef oneindig vaak opnieuw trokken. Merk op: hiervoor heb je geen nauwkeurige (efficiënte) schatter nodig!
- Een schatter is *efficiënt* indien zijn steekproevenverdeling een kleine standaardfout heeft (hij lijkt dan meestal sterk op de parameter).

Bij aselecte steekproeftrekking zijn schatters niet per se efficiënt (dat hangt af van de steekproefomvang), maar wel vaak zuiver, zolang de juiste schatter voor de juiste parameter wordt gebruikt:
- $\overline{X} \rightarrow \mu$
- $b \rightarrow \beta$

Enzovoort.

Simpel effect
Zie *Interactie, hoofdeffecten en simpele effecten*

Steekproevenverdeling *Geïntroduceerd in boek 1, hoofdstuk 5*
Een toevalsvariabele, zoals een *schatter* (zie eerder), kan elke keer een andere waarde aannemen als we een nieuwe steekproef trekken. We kunnen daarom een kansverdeling opstellen van alle mogelijke waarden die deze toevalsvariabele kan aannemen over steekproeven heen. Die kansverdeling noemen we de steekproevenverdeling van de variabele.
De belangrijkste parameters van de meeste steekproevenverdelingen zijn hun gemiddelde en standaardafwijking:
- De *verwachtingswaarde*. Dit is de gemiddelde waarde die de toevalsvariabele zou aannemen, als we een oneindig aantal steekproeven trokken. Bij aselecte steekproeftrekking is deze verwachtingswaarde vaak gelijk aan een bepaalde populatieparameter. Bijvoorbeeld: het gemiddelde van alle mogelijke steekproefgemiddelden die je zou kunnen trekken, zal altijd gelijk zijn aan het gemiddelde van de populatie. We zeggen dan dat het gemiddelde van de steekproef een *zuivere schatter* is van het populatiegemiddelde.
- De *standaardfout*. Dit is de gemiddelde afwijking van de toevalsvariabele van diens gemiddelde (zijn verwachtingswaarde). De standaardfout vertelt ons dus hoe sterk de steekproefschatter doorgaans de populatieparameter over- of onderschat (gegeven dat-ie zuiver is). Anders dan de verwachtingswaarde kun je de standaardfout beïnvloeden door een grotere steekproef te trekken; dit maakt de standaardfout kleiner. Als de standaardfout klein is, zeggen we dat de steekproefschatter (zoals een gemiddelde) een *efficiënte schatter* is van de populatieparameter die we zoeken.

Tot slot zullen vele steekproevenverdelingen ongeveer *normaal* zijn van vorm als de getrokken steekproef *groot genoeg* is, ongeacht hoe de populatie eruitziet; een wet die bekendstaat als de *centrale-limietstelling*.

Sum of squares	Toevalsvariabele
Zie *Variantie*	Zie *Steekproevenverdeling*

Type I-fout, type II-fout en power *Geïntroduceerd in boek 1, hoofdstuk 6*
Bij hypothesetoetsen (zie elders) komt onzekerheid kijken, dus ze kunnen goedgaan of foutgaan. Hier heb je een overzicht van wat er kan gebeuren (tabel 0.9):

TABEL 0.9 Mogelijke uitslagen van een hypothesetoets

		Toetsresultaat	
		H_0 niet verworpen	H_0 verworpen
Werkelijkheid	H_0 waar	Juiste beslissing $1 - \alpha$	Type I-fout α
	H_0 onwaar	Type II-fout β	Juiste beslissing *power* = $1 - \beta$

Dus:
- Je maakt een *type I-fout* als je een ware nulhypothese onterecht verwerpt. De kans dat dit gebeurt (bij een enkele toets) is altijd gelijk aan je *significantieniveau* α.
- Je maakt een *type II-fout* als het je niet lukt een onware nulhypothese te verwerpen.
- De *power* is de kans dat je erin zult slagen een onware nulhypothese te verwerpen.

Hoe groter de power, hoe beter. We kunnen die op diverse manieren vergroten, met name door:
- de steekproef te vergroten;
- de errorterm of residuele term van het statistische model te verkleinen.

Hierover meer in de betreffende hoofdstukken. De power hangt ook af van:
- het significantieniveau (maar dat kun je beter niet verhogen voor meer power, want die tactiek leidt tot meer type I-fouten);
- de effectgrootte (maar die laat zich vaak niet vergroten voor meer power).

Om de power van een z-toets gemakkelijk met de hand te bepalen, bereken je zijn z-score:

$$z_{1-\beta} = z^* - \frac{|\mu_{werkelijk} - \mu_0|}{\frac{\sigma}{\sqrt{N}}}$$

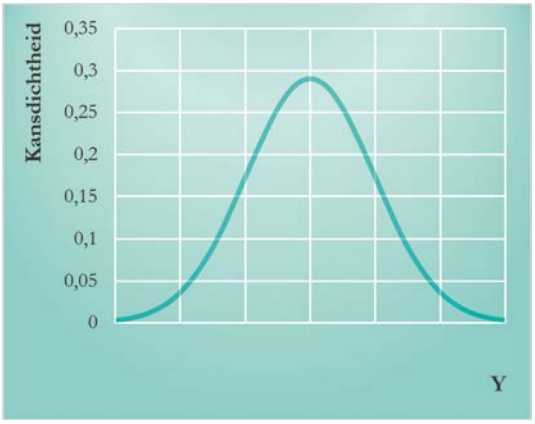
Vervolgens zoek je de *rechtszijdige* kans op deze z-score op in de kanstabel. Meer details vind je in het hoofdstuk Poweranalyse in de online leeromgeving.

Variatie
Zie *Variantie*

Variantie *Geïntroduceerd in boek 1, hoofdstuk 1 en 10*

Wat variabelen definieert is dat mensen er verschillend op gescoord hebben. Vanwege dit feit kennen we score*verdelingen* (figuur 0.7). Is jouw variabele kwantitatief, dan kun je de spreiding in de scores onder meer uitdrukken met een *standaardafwijking* of *standaarddeviatie*: de gemiddelde afwijking van het gemiddelde. De berekening bestaat uit drie stappen, waarvan de tweede de variantie oplevert.

FIGUUR 0.7 Spreiding



1 Bepaal hoe ver iedere respondent afwijkt van het gemiddelde.

$$Y_{ij} - \mu$$

Om de gemiddelde afwijking te bepalen, moet je eerst alle afwijkingen optellen. Ze zullen elkaar opheffen, tenzij je ze kwadrateert. Tel dus in plaats hiervan de gekwadrateerde afwijkingen op, hetgeen je de *kwadratensom* oplevert (Engels: *sum of squares*):

$$kwadratensom\ (SS) = \sum_{ij}(Y_{ij} - \mu)^2$$

Je mag de kwadratensom ook wel de *variatie* in de scores noemen (deze term is in het Engels niet heel gangbaar, maar in het Nederlands wel).

2 Pak de *gemiddelde* gekwadrateerde afwijking per persoon. Voor populaties houdt dit in dat je de kwadratensom deelt door N, maar voor steekproeven moet je de kwadratensom delen door zijn *vrijheidsgraden*, $N - 1$. Het resultaat noemen we de *variantie* (in het Engels ook wel *mean square*):

$$variantie = \frac{kwadratensom\ (SS)}{vrijheidsgraden\ (df)}$$

3 Trek de wortel om van het kwadraat af te komen. De waarde die eruitkomt laat zich gemakkelijker interpreteren, want hij staat voor de gemiddelde afwijking van het gemiddelde. De *standaardafwijking* is dus

$$standaardafwijking = \sqrt{variantie} = \sqrt{\frac{kwadratensom\ (SS)}{vrijheidsgraden\ (df)}}$$

Kortweg zijn varianties gemakkelijker om mee te rekenen (geen wortel), maar standaardafwijkingen zijn gemakkelijker om uit te leggen.
In ANOVA- en regressiemodellen werken we meerdere kwadratensommen uit. Die delen we vervolgens door hun eigen soort vrijheidsgraden. De vari-

anties die hier uitkomen, worden in het Engels meestal *mean squares* genoemd.

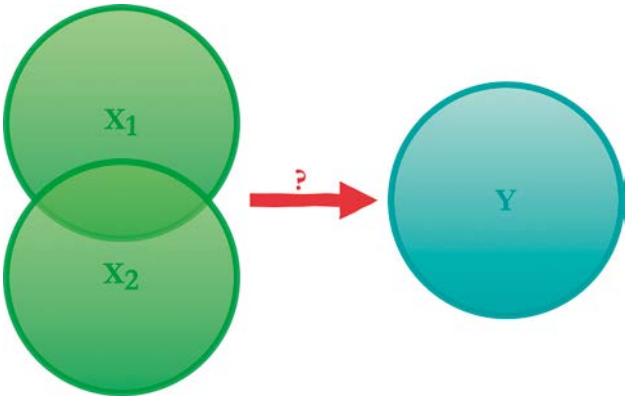
'Kijk, ik heb een *mean square* getekend!'
~ Simeon De Smet

Verstoring *Geïntroduceerd in hoofdstuk 3, 8 en 10*

Dit is een *meetprobleem*. Het doet zich voor als we niet het 'echte' effect meten dat een bepaalde onafhankelijke variabele heeft, doordat een andere onafhankelijke variabele onze observatie vertekent. Zo ontdekte Granger in hoofdstuk 8 dat volbloed tovenaars (van wie beide ouders magie konden gebruiken) vaker overliepen naar de Zwarte Zijde. Maar toen ze haar data nader bestudeerde (zoals in de paragraaf over *orthogonaliteit*), zag ze dat de volbloed magiërs vaak sociaal conservatief waren – veel meer dan magiërs met gemengd bloed. Met andere woorden, bloedstatus en conservatisme overlapten elkaar deels (ze vertoonden een samenhang), waardoor het moeilijker werd om vast te stellen welke variabele *werkelijk* verantwoordelijk was voor de zijde waaraan de tovenaars trouw zwoeren (de Witte of de Zwarte). Als een onderzoeker gelooft (op basis van zijn of haar analyse) dat het bloedstatus was, maar het in feite om conservatisme ging, is die laatste variabele een *verstoorder* (figuur 0.8).

Je moet altijd proberen verstoring te corrigeren en het 'pure' effect te meten dat elke variabele heeft. Meerweg-ANOVA-modellen en meervoudige regressiemodellen doen dit vanzelf, mits je de potentiële verstoorders opneemt in je analyse.

FIGUUR 0.8 Verstoring



Drie verdere opmerkingen:

- Een verstoorder kan ook het geobserveerde effect van een andere onafhankelijke variabele *onderdrukken*. Dit houdt in dat die onafhankelijke variabele wel een effect heeft, maar de verstoorder doet dit effect zwakker lijken of maakt het zelfs onzichtbaar. Stel je voor dat X_1 een positief effect heeft op de afhankelijke variabele Y, maar X_2 heeft een negatief effect; dan kunnen ze tegen elkaar in werken. (Dit gebeurt wanneer X_1 tevens een positieve correlatie heeft met X_2.)
- Verstoring vereist overlap tussen de onafhankelijke variabelen, anders kan ze niet optreden. Als zodanig is ze alleen mogelijk in *non-orthogonale* designs.
- Verstoring heeft *niets* te maken met interactie (zie een eerdere paragraaf). Interactie kan ook bestaan als een echt effect in de populatie, en ze doet zich voor in zowel orthogonale als non-orthogonale designs. Ze houdt in dat sommige x-variabelen *daadwerkelijk* elkaars effecten op de afhankelijke variabele Y veranderen. Merk op hoe interactie altijd vanaf het begin de afhankelijke variabele erbij betrekt, terwijl het risico op verstoring verrijst uit overlap tussen de *onafhankelijke* variabelen.

Vrijheidsgraden *Geïntroduceerd in boek 1, hoofdstuk 7*

Ik ben erg gelukkig met mijn uitleg van vrijheidsgraden in de online leeromgeving, al zeg ik het zelf. Hij is simpel en compleet. ☺ Ik geef je liever geen verwaterde samenvatting, dus ga maar even naar *statistiekinstijl.noordhoff.nl*. Heb je toch graag een snellere indruk? Tot je dienst:

- Vrijheidsgraden zijn het aantal onafhankelijke observaties die we kunnen doen wanneer we een bepaalde steekproefschatter berekenen, of het gedrag bestuderen van een bepaalde variabele. Laten we zeggen dat je een dek met 30 rode en zwarte kaarten hebt, en je begint de rode te tellen. Je komt uit op 18 rode kaarten; dit wist je niet van tevoren. Maar nu weet je wél meteen dat het aantal zwarte kaarten gelijk moet zijn aan ... 12. Zodra je dus de ene kleur hebt geteld, staat het aantal kaarten van de andere kleur vast. De kleurvariabele, zeggen we dan, heeft slechts één vrijheidsgraad.

- Meer vrijheidsgraden geven de schatter of variabele meer 'bewegingsruimte': hij kan meer veranderen als we een nieuwe steekproef trekken. Stel je maar voor dat we een derde kaartkleur in de stapel doen.
- Er zijn verschillende soorten vrijheidsgraden. Sommige hangen af van de omvang van de steekproef (N). Dit zijn altijd de vrijheidsgraden in de zogenoemde 'errorterm' (ANOVA) of 'residuele term' (regressie). Meer vrijheidsgraden van dit type geven *meer power*.

Z-scores *Geïntroduceerd in boek 1, hoofdstuk 1 en 5*

De z-score is een veelgebruikt soort *datatransformatie* (zie elders) voor kwantitatieve variabelen. De truc van z is dat we de schaal een gemiddelde geven van 0 en een standaardafwijking van 1. Ziehier de algemene formule:

$$z = \frac{X - \mu}{\sigma}$$

Oftewel, x-score *(observatie)* minus gemiddelde *(verwachting)*, gedeeld door de standaardafwijking. De formule kan aangepast worden voor verschillende soorten verdelingen. Voor steekproeven gebruiken we $z = \frac{X - \overline{X}}{s}$, en voor steekproevenverdelingen van het gemiddelde hebben we $z = \frac{\overline{X} - \mu}{\left(\frac{\sigma}{\sqrt{N}}\right)}$.

Zuivere schatter
Zie *Schatter*

DEEL 1
Data beschrijven

1 Dataverkenning voor gevorderden 35

Welkom bij *Statistiek in stijl 2*! Statistiek staat op het punt een serieuze zaak te worden. Dit eerste deel bevat één hoofdstuk waarin je kennismaakt met professionelere methoden om een steekproef te verkennen. Vrienden worden met je data is een essentiële stap in het onderzoeksproces. Hoe beter je begrijpt wat je ziet – een koppige kat, abnormale meerminnen of smeltende ijsjes – des te sterker worden je conclusies over de populatie. Laat het volgende niveau beginnen!

1 Dataverkenning voor gevorderden

Over zeemeerminnen en zuivere schatters

Niveau
Gorilla
Leerdoelen
Na dit hoofdstuk kun je Beargumenteren of een vreemde score een uitschieter is met grote invloed Adequate schatters voor effectgrootten gebruiken Genuanceerd naar assumpties kijken van statistische modellen Variabelen monotoon transformeren

1.1 Een uitschieter zijn of niet zijn (dat is de vraag) 36
1.2 Klein maar fijn: effectgrootten schatten 42
1.3 Zien is geloven? Assumpties opnieuw bekeken 49
1.4 Metamorfose: monotone transformaties 60
Samenvatting 68
1A Opdrachten 69

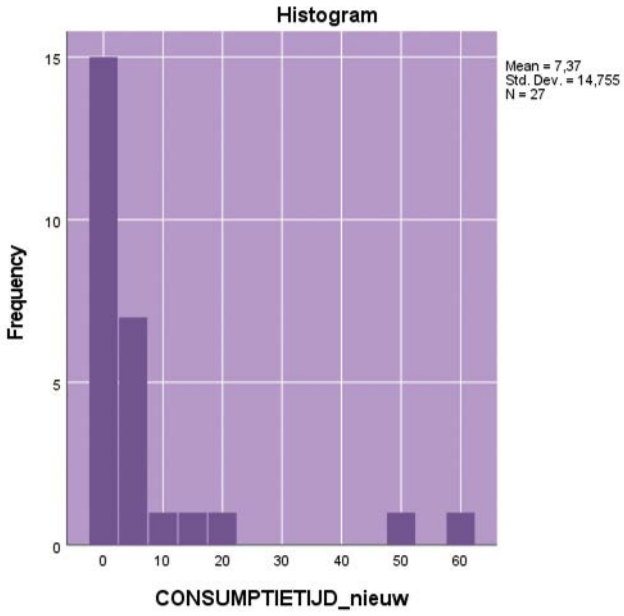
1.1 Een uitschieter zijn of niet zijn (dat is de vraag)

Vereiste voorkennis	Handig om te hebben
Boek 1, hoofdstuk 1: Gegevens in kaart	Boek 1, hoofdstuk 6: Hypothesetoetsing

Termen met een *groene markering* vind je in *Overzichten en kernbegrippen* vooraan in dit boek.

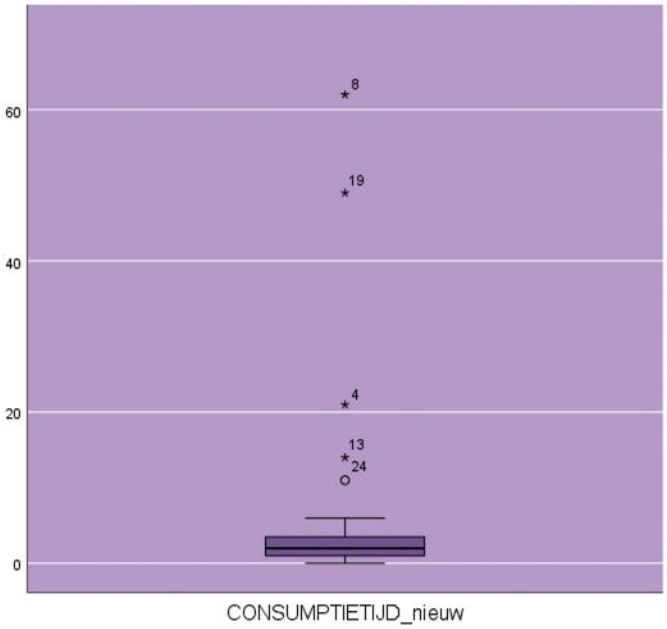
In hoofdstuk 8 uit *Statistiek in stijl 1* verbeterde de Duitse fabrikant Witz-Katz zijn kattenvoerrecept en liet hij een steekproef van $N = 27$ katten het recept proberen. De variabele waar men naar keek, was hoelang de kat erover deed om één schaaltje *Fischmasch*® leeg te eten. Hoe korter dit duurde, des te meer moet het gesmaakt hebben. Nietwaar?

FIGUUR 1.1 Histogram van de consumptietijd



Hier heb je een histogram van de data voor het nieuwe recept (figuur 1.1). Het lijkt een paar flinke uitschieters te bevatten. Het boxplot (figuur 1.2) maakt dit nog duidelijker: uitschieters staan buiten de box geplaatst en aangeduid met cirkels of sterren. Het ziet er dus naar uit dat meerdere katten in hongerstaking zijn gegaan toen ze Fischmasch kregen; proefkat nummer 8 weigerde meer dan 60 uur lang het voer op te eten. Wat voor een criterium heeft SPSS hier echter gebruikt? En welke andere criteria zijn er om scores als uitschieters te bestempelen?
In boek 1 hebben we een enkel criterium besproken. Tijd om ons arsenaal uit te breiden. Statistiek is geen exacte wetenschap, ondanks wat ons onderbuikgevoel ons vertelt als we getalletjes zien. Verschillende criteria zijn het niet altijd eens, dus het is verstandig om er meerdere te bekijken en een genuanceerd oordeel te vormen.

FIGUUR 1.2 Boxplot van de consumptietijd



Histogram: op zoek naar eilandjes

+ Beresimpel.
– Een beetje subjectief.
– Vertelt je niet hoeveel invloed de uitschieter heeft.

Zie je scores die 'eilandjes' vormen, los van de grote 'landmassa'? Dat kunnen uitschieters zijn. Merk op dat de gaten het algemene patroon moeten doorbreken. Figuur 1.1 bevat inderdaad ver verwijderde eilandjes.

Lijsten van extreme waarden

+ Beresimpel.
– Een beetje subjectief.
– Vertelt je niet hoeveel invloed de uitschieter heeft.

Zoals je ziet, heeft dit criterium dezelfde plussen en minnen als een inspectie van het histogram. Dat komt doordat het in wezen dezelfde aanpak is, alleen kijken we naar een aantal preciezere waarden. We kunnen SPSS vragen een overzicht te maken van de vijf hoogste scores op een rij, en de vijf laagste:

Wil je meedoen? Ga dan naar *statistiekinstijl.noordhoff.nl*, log in bij de online leeromgeving en download de instructies en gegevensbestanden voor hoofdstuk 1.

1

Extreme Values

			Case Number	Value
CONSUMPTIETIJD nieuw	Highest	1	8	62
		2	19	49
		3	4	21
		4	13	14
		5	24	11
	Lowest	1	12	0
		2	9	0
		3	1	0
		4	25	1
		5	23	1[a]

a. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

Laten we de lijst met hoge extremen naslaan, aangezien we enkele opwaartse uitschieters verwachten (op basis van het vorige criterium). Is de hoogste score echt zo ver verwijderd van de rest, of volgen de afstanden tussen de scores een redelijk natuurlijk verloop? Hier lijkt het eerste het geval. We hebben een kat die de Fischmasch in 11 uur heeft opgegeten, gevolgd door eentje die er 14 uur over deed, dan een score van 21 … en vervolgens een enorme sprong naar 49 uur. De laatste score is zelfs 62; nog een grote sprong vanaf 49. Dit duidt er sterk op dat de hoogste twee scores uitschieters zijn. Ze behoren tot kat nummer 8 en 19 (zie het *Case Number*) in de dataset van WitzKatz.

IQR-criterium

\+ Gebaseerd op een rekenregel.
– Werkt minder goed voor scheve verdelingen.

Interkwartielafstand

Kwartiel

Boek 1 legt de *interkwartielafstand (IQR)* helemaal uit, maar ik zal hem hier kort samenvatten. De box in figuur 1.2 bevat 50% van de scores, dus als je goed kijkt, zie je dat 50% van de katten in de steekproef tussen de 1 en 4 uur deed over het eten van de Fischmasch. 1 en 4 zijn daarom het *eerste* en *derde kwartiel* (Q_1 en Q_3) van de scoreverdeling. (Het tweede kwartiel is de *mediaan*, de dikke lijn in het midden van de box; die is gelijk aan 2.) Wat

was dus de afstand tussen de middelste 50%? Ze lagen hooguit 4 − 1 = 3 uur uit elkaar. Dit is de IQR:

$$IQR = Q_3 - Q_1$$

De *Descriptives*-tabel hierna geeft hem eveneens weer (deze kunnen we verkrijgen met dezelfde SPSS-analyse als de *Extreme Values*-tabel). Hetgeen ons de moeite van een handmatige berekening bespaart.

Descriptives

			Statistic	Std. Error
CONSUMPTIETIJD nieuw	Mean		7,37	2,840
	95% Confidence Interval for Mean	Lower Bound	1,53	
		Upper Bound	13,21	
	5% Trimmed Mean		4,93	
	Median		2,00	
	Variance		217,704	
	Std. Deviation		14,755	
	Minimum		0	
	Maximum		62	
	Range		62	
	Interquartile Range		3	
	Skewness		3,024	,448
	Kurtosis		8,863	,872

1,5 × IQR-criterium

Om het zogenoemde *1,5 × IQR-criterium* toe te passen, berekenen we twee grenzen (let op de subtiele verschillen in de twee formules):

$$Q_1 - (1{,}5 \times IQR) = 1 - (1{,}5 \times 3) = 1 - 4{,}5 = -3{,}5$$
$$Q_3 + (1{,}5 \times IQR) = 4 + (1{,}5 \times 3) = 4 + 4{,}5 = 8{,}5$$

De gedachte is dat we alle scores tussen –3,5 en 8,5 als acceptabel kunnen beschouwen … maar dat scores die *buiten dit bereik* vallen uitschieters zijn. Dit verklaart waarom figuur 1.2 de katten met een consumptietijd boven de 8,5 uur uit de box heeft verjaagd.
Uitschieters volgens het voorgaande criterium staan gemarkeerd met cirkeltjes. Maar SPSS gebruikt sterretjes voor de exemplaren die het *3 × IQR-criterium* schenden:

3 × IQR-criterium

$$Q_1 - (3 \times IQR) = 1 - (3 \times 3) = 1 - 9 = -8$$
$$Q_3 + (3 \times IQR) = 4 + (3 \times 3) = 4 + 9 = 13$$

Asterisken duiden daarom op *extreme* uitschieters, die je aandacht beslist waard zijn. Ze maken meer kans om ook de sirenes van de andere criteria te laten afgaan.

5% getrimd gemiddelde

+ Kan suggereren hoe invloedrijk de uitschieter is.
– Een beetje subjectief.

Werp nogmaals een blik op de *Descriptives*-tabel hiervoor. Het 5% getrimd gemiddelde *(5% trimmed mean)* vertelt ons wat de gemiddelde consumptietijd zou zijn, als we de 5% meest extreme scores wegknipten. Dat is 2,5% aan de onderkant en 2,5% aan de bovenkant. Als het getrimde gemiddelde flink verschilt van het gewone gemiddelde, laat dit zien dat minstens één uitschieter hier een sterke invloed op heeft. Hetgeen hier het geval lijkt: als we een paar katten trimmen, zakt het gemiddelde van 7,37 naar 4,93 uur.

Z-scores

+ Gebaseerd op een rekenregel.
– Werkt minder goed voor non-normale verdelingen.

In een normale verdeling ligt ongeveer 95% van de scores binnen 2 standaardafwijkingen van het gemiddelde. Anders gezegd, deze scores hebben een z-score tussen de –2 en 2. Als extremere waarden vaker dan 5% van de tijd opduiken, zijn het mogelijk uitschieters. Regel dus een simpele frequentietabel en begin te tellen:

Zscore(CONSUMPTIETIJD nieuw)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-,49952	3	11,1	11,1	11,1
	-,43175	7	25,9	25,9	37,0
	-,36397	5	18,5	18,5	55,6
	-,29620	5	18,5	18,5	74,1
	-,22843	1	3,7	3,7	77,8
	-,09288	1	3,7	3,7	81,5
	,24600	1	3,7	3,7	85,2
	,44932	1	3,7	3,7	88,9
	,92374	1	3,7	3,7	92,6
	2,82143	1	3,7	3,7	96,3
	3,70250	1	3,7	3,7	100,0
	Total	27	100,0	100,0	

We hebben twee z-waardes boven de 2, elk met een frequentie van 1. Dat is $\frac{2}{27} \triangleq 7{,}4\%$ van de totale steekproefomvang. De WitzKatz-gegevens bevatten dus een groter aantal extreme z-scores dan je zou verwachten op basis van een normale verdeling (5%). Deze extreme z-scores lijken uitschieters.

'Maar Vincenzo, de consumptietijd was totaal niet normaal verdeeld. Hoeveel hebben we dan aan deze regel?' Goed punt, waarde lezer. In zo'n geval zou ik de grenzen voor de z-scores oprekken tot *–2,5 en 2,5*. Zelfs dan zien de uitschietende katten er extreem uit. Een liberalere richtlijn zou *–3 en 3* zijn; de aanbeveling verschilt een beetje afhankelijk van wie je het vraagt.

Omgaan met uitschieters

Niet alle uitschieters zijn problematisch. Iedereen heeft het recht om raar te zijn. Zolang scores je statistieken niet verstoren, kun je ze zonder aanpassingen in je dataset laten zitten. Maar als ze wel tot vertekening leiden, wat doen we dan? De beste oplossing hangt misschien af van de situatie. Hier komen een paar populaire suggesties:

De uitschieter(s) weghalen
+ Simpel.
– Moet theoretisch verdedigbaar zijn.
– Je steekproef wordt kleiner, hetgeen de *power* kan aantasten.

Je mag een kat alleen maar het raam uit gooien als je kunt aantonen dat-ie niet representatief is voor de populatie die je onderzoekt. In dit geval is het echter heel typisch voor katten om onbuigzaam te zijn en botweg te weigeren onsmakelijk voer te eten … nietwaar? Het lijkt dan ook ongepast om ze gewoon weg te laten en net te doen alsof deze extreme poezen niet bestaan.
Wanneer mogen we uitschieters wél verwijderen? Stel dat de Fischmasch zich qua marketing richtte op kittens, en stel dat de uitschieters per ongeluk norse oude katers bleken te zijn. In zo'n geval zouden ze niet de kattenpopulatie vertegenwoordigen die de fabrikant interesseerde. Voel je dan niet bezwaard en haal ze gewoon uit de dataset.

Trimmen
+ Een theoretische onderbouwing is niet nodig.
– Verwijdert vaak ook gewone scores.
– Je steekproef wordt kleiner, hetgeen de *power* kan aantasten.

Afkappen

Ook wel *afkappen* genoemd (*truncating* in het Engels). In feite was ook de vorige methode een vorm van trimmen, maar deze benadering houdt in dat we zonder onderscheid de laagste en de hoogste scores wegscheren. Bijvoorbeeld: een 5%-trim betekent dat we 2,5% aan de onderkant verwijderen en 2,5% aan de bovenkant. Dit maakt het steekproefgemiddelde robuuster tegen uitschieters (zie ook het 5% getrimde gemiddelde) en het verkleint de standaardafwijking. SPSS biedt vooralsnog geen procedure om de scores te trimmen, dus dit kun je enkel met de hand doen.
(Eén tip: pas ontzettend op met het trimmen van een kat, vooral als die boos en hongerig is …)

Winsoriseren

+ Verwijdert geen scores.
– Wijzigt oorspronkelijke scores.

Deze methode schakelt de uitschietersscore gelijk aan de meest extreme score die *geen* uitschieter is. Neem als voorbeeld nogmaals de *Extreme Values*-tabel van eerder. Laten we zeggen dat we de op twee na hoogste score (21) nog als een gewone beschouwen. In dat geval stelde Charles Winsor voor dat we de uitschieters hiervoor (49 en 62) eveneens in 21 veranderen. Gebruiken we SPSS, dan zullen we dat handmatig moeten doen.

Datatransformatie

+ Verwijdert geen scores.
– Vermindert de impact van uitschieters, maar rekent er niet altijd mee af.
– Maakt je variabele moeilijker te interpreteren, dus dit is vooral geschikt voor het berekenen van geldige hypothesetoetsen en betrouwbaarheidsintervallen. Probeer de transformatie terug te draaien zodra je deze hebt.

Paragraaf 1.4 bespreekt deze methode in detail.

1.2 Klein maar fijn: effectgrootten schatten

Vereiste voorkennis

Boek 1, hoofdstuk 9: Ongepaarde t-toets
Boek 1, hoofdstuk 10: Eenweg-ANOVA

Een significant effect zegt niet alles. De grootte ervan is minstens even belangrijk. Maar hoe kunnen we effectgrootten uitdrukken op een toegankelijke manier en ze tegelijk zo zuiver mogelijk schatten? Deze paragraaf vist het uit in detail.

I Globale effectmaat: η^2

In hoofdstuk 10 van *Statistiek in stijl 1* bestudeerde onderzoeker Andersen $N = 87$ meerminnen die zich onzeker voelden over hun uiterlijk, en dit wilden compenseren met een vinvergroting. Eén derde van hen kreeg de chirurgie waar ze om gevraagd hadden, één derde nam in plaats daarvan groepstherapie om een immaterialistische levenshouding te ontwikkelen, en de rest werd gedurende het experiment op een wachtlijst geplaatst. Twee maanden later mat Andersen met een vragenlijst hoe tevreden ze waren met hun lichaam.

Willen we een hele populatie beschrijven, dan doet een eenweg-ANOVA dit als volgt:

$$\sigma_Y^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$$

In woorden:
- De mate waarin alle meerminnen qua lichaamsbeeld verschilden (de *totale variantie* σ_Y^2), werd bepaald door twee dingen ...
- ... namelijk de *behandelingsvariantie* σ_α^2 (de variantie van de *populatiegemiddelden* van de diverse groepen) ...
- ... en de *errorvariantie* σ_ε^2 (de variantie tussen de *individuen* in dezelfde populatie). Het ANOVA-model gaat er hierbij van uit dat die errorvariantie in elke populatie even groot is (dit is een van de assumpties).

Proportie verklaarde variantie

Hoe groter het effect van de behandeling, des te meer bestaat de totale variantie uit behandelingsvariantie. Voor hoeveel? Dat drukken we uit met η^2 (èta-kwadraat): de *proportie verklaarde variantie*.

$$\eta^2 = \frac{\sigma_\alpha^2}{\sigma_Y^2}$$

Is η^2 bijvoorbeeld gelijk aan 0,10, dan wordt de totale spreiding in de lichaamsbeeldscores voor 10 procent verklaard door het feit dat de meerminnen verschillende behandelingen kregen. Merk op dat we hier spreken over *populatieparameters*. De formule die je in boek 1 hebt gezien, vormt een specifieke *steekproefschatter* voor de η^2 in de populatie! Hier komen we zo dus op terug. Maar eerst volgt hier nogmaals de richtlijn die Jacob Cohen formuleerde voor η^2:
- 0,01: klein effect;
- 0,06: medium effect;
- 0,14: groot effect.

Dit is natuurlijk maar een vuistregel. Bovendien geldt-ie niet universeel: in sommige vakgebieden kan 0,10 al een zeldzaam groot effect zijn. In zo'n geval kun je beter kijken of er voor jouw veld geen specifiekere richtlijnen zijn ontwikkeld.
Dan nu de vraag: hoe kunnen we η^2 het beste schatten op basis van Andersens steekproef? Hiervoor zijn in de loop der jaren maar liefst drie steekproefschatters ontwikkeld:
- $\hat{\eta}^2$. Let op het hoedje, dat 'schatting' betekent. Deze is in boek 1 aan bod gekomen. Helaas is hij, zo blijkt, geen *zuivere schatter* van η^2;
- $\hat{\omega}^2$ (omega-kwadraat). Deze populaire variant is zuiverder dan $\hat{\eta}^2$;
- $\hat{\varepsilon}^2$ (epsilon-kwadraat). In tegenstelling tot wat veel statistici denken, lijkt deze de zuiverste schatter te zijn!

Als rekenvoorbeeld volgt hierna Andersens ANOVA-tabel. We gaan de schatters nu een voor een bestuderen.

TABEL 1.1 ANOVA-tabel bij Andersens data

	Kwadratensom *(sum of squares)*	Vrijheidsgraden *(df)*	Variantie *(mean square)*	F-ratio
Tussen groepen *(between groups, G)*	1180,023	2	590,011	10,748
Binnen groepen *(within groups, E)*	4611,379	84	54,897	
Totaal *(total, T)*	5791,402	86	(67,342)	

$\hat{\eta}^2$ (geschatte èta-kwadraat)

Deze statistiek heeft de simpelste formule. Hij pakt gewoon de kwadratensommen *(sums of squares)* en bekijkt hoeveel van de totale variatie bestaat uit variatie tussen de groepen:

$$\hat{\eta}^2 = \frac{SS_{between}}{SS_{total}} = \frac{1180{,}023}{5791{,}402} = 0{,}204$$

Volgens $\hat{\eta}^2$ verklaart de behandeling dus voor 20,4% de variatie in de lichaamsbeelden.
Wat mankeert er aan deze schatter? Nou, hij zit meestal te hoog; hij overschat de echte η^2 dus systematisch. De voornaamste reden is dat verschillende steekproeven door toeval vrijwel altijd wel een beetje zullen verschillen, zelfs als de populatiegemiddelden exact hetzelfde zijn. $\hat{\eta}^2$ wordt dan groter dan 0 (overschatting), maar kan überhaupt nooit kleiner dan 0 worden (onderschatting). Die onder- en overschattingen heffen elkaar gemiddeld dus niet mooi op.

$\hat{\varepsilon}^2$ (epsilon-kwadraat)

Je zou kunnen zeggen dat $SS_{between}$ vaak te groot is. Hij beschrijft de variatie tussen de groepsgemiddelden, maar die variatie kan ook door restfactoren komen (error). $\hat{\varepsilon}^2$ filtert die errorinvloeden zo goed mogelijk weg uit de effectschatting. Namelijk zo:

$$\hat{\varepsilon}^2 = \frac{SS_{between} - df_{between}MS_{within}}{SS_{total}} = \frac{1180{,}023 - 2 \times 54{,}897}{5791{,}402} = \frac{1070{,}229}{5791{,}402} = 0{,}185$$

De precieze logica achter deze formule is een beetje technisch, maar zoals je ziet wordt de teller kleiner in vergelijking met $\hat{\eta}^2$. MS_{within} is een zuivere schatter van de errorvariantie (zie boek 1) en die error willen we weghalen uit de teller van de breuk.

$\hat{\omega}^2$ (omega-kwadraat)

Volgens de aanpak van Hays (1964) is niet alleen $SS_{between}$ vaak te groot, maar ook is SS_{total} vaak te klein. Dit is wederom een beetje technisch om

precies uit te leggen, maar bij wijze van correctie voegde Hays nog eens MS_{within} toe aan SS_{total}. Het resultaat noemen we $\hat{\omega}^2$:

$$\hat{\omega}^2 = \frac{SS_{between} - df_{between}MS_{within}}{SS_{total} + MS_{within}} = \frac{1180{,}023 - 2 \times 54{,}897}{5791{,}402 + 54{,}897} = \frac{1070{,}229}{5846{,}299} = 0{,}183$$

Het is trouwens niet zo dat $\hat{\omega}^2$ een poging is om $\hat{\varepsilon}^2$ te 'verbeteren'; verschillende statistici hebben deze twee schatters onafhankelijk van elkaar uitgevonden.

Welke moet ik gebruiken?
$\hat{\eta}^2$ zou ik vanaf nu niet meer nemen; vooral bij kleine steekproeven kan die de echte effectgrootte flink overschatten. De gangbare opvatting is dat $\hat{\omega}^2$ de zuiverste schatter is. Die opvatting lijkt echter gebaseerd op een computersimulatie van Keselman uit 1975. Okada (2013) en Lakens (2015) hebben nieuwe simulaties uitgevoerd, die erop wijzen dat $\hat{\varepsilon}^2$ het zuiverst is. Ik vind hun artikelen overtuigend. ☺ Dus als je het van mij moet hebben, raad ik $\hat{\varepsilon}^2$ aan. Maar ook $\hat{\omega}^2$ is lang niet slecht.
Voer je met SPSS een eenweg-ANOVA uit, dan kun je vanaf versie 27 (eindelijk) al deze effectmaten laten uitrekenen door het programma. Je hoeft daarvoor in het *Oneway*-venster enkel te klikken op het extra knopje *Estimate effect sizes for overall tests*. Deze tabel komt er dan bij:

ANOVA Effect Sizes[a]

			95% Confidence Interval	
		Point Estimate	Lower	Upper
LICHAAMSBEELD	Eta-squared	,204	,062	,335
	Epsilon-squared	,185	,039	,319
	Omega-squared Fixed-effect	,183	,039	,316
	Omega-squared Random-effect	,101	,020	,188

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

Fraai is dat je niet alleen de steekproefschatter krijgt *(Point Estimate)*, maar ook een *betrouwbaarheidsinterval*. Zo kunnen we volgens èta-kwadraat 95% zeker zijn dat η^2 in de populatie tussen 0,062 en 0,335 ligt. Epsilon-kwadraat is (terecht) wat pessimistischer en houdt het op een η^2 (zonder hoedje) tussen 0,039 en 0,319.

II Lokale effectmaat: δ

Als we slechts 2 groepen vergelijken, is het heel simpel om een effectmaat te bepalen. We bekijken gewoon hoe de groepen verschillen qua *gemiddelde*, en delen dit door de standaardafwijking van de individuele scores (binnen een groep):

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Zo maken we van het verschil een *z-score*. σ kan de standaardafwijking van beide populaties zijn (aangenomen dat ze dezelfde hebben) of van een van de populaties; hierover verderop meer. Jacob Cohen heeft voor δ (delta) deze richtlijn opgesteld:

- 0,2: klein effect;
- 0,5: medium effect;
- 0,8: groot effect.

Sawilowski (2009) heeft dit nog uitgebreid met 1,2 voor een 'heel groot effect' en 2,0 voor een 'enorm effect'. Hoe moeten we δ schatten als we enkel een steekproef trekken? Hier volgen de data uit Andersens onderzoek:

Descriptives

LICHAAMSBEELD

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
vinvergroting	29	17,45	11,262	2,091	13,16	21,73	0	36
groepstherapie	29	12,41	3,942	,732	10,91	13,91	6	20
controle	29	8,45	4,725	,877	6,65	10,25	1	17
Total	87	12,77	8,206	,880	11,02	14,52	0	36

Natuurlijk kan δ maar met twee groepen tegelijk. We kunnen er dus een bepalen voor het verschil tussen de vinvergroting en de controlegroep; een voor het verschil tussen groepstherapie en controle; en een voor het verschil tussen vinvergroting en groepstherapie. Vandaar dat ik δ een *lokale* effectmaat noem: als je meer dan twee groepen hebt, zoomt hij steeds in op één paar.

Cohens *d* en Hedges' *g*

Die eerste is een oude bekende uit boek 1. We berekenen hem zo:

$$d = \frac{|\overline{Y}_1 - \overline{Y}_2|}{s_p}$$

Voor de hand liggend dus: we nemen het (absolute) verschil tussen twee steekproefgemiddelden en delen dit door de *gepoolde* standaardafwijking. Die laatste is een gewogen gemiddelde van alle standaardafwijkingen. Over de precieze formule zijn verschillende statistici het niet eens, dus laten we die discussie maar uit de weg gaan. Belangrijker is dat een gemiddelde van de standaardafwijkingen alleen ergens op slaat als alle populaties *dezelfde* standaardafwijking hebben; anders gezegd, als er aan de assumptie van gelijke varianties is voldaan (zie paragraaf 1.3). Gebruiken we de aanpak uit boek 1, dan kunnen we stellen dat s_p gelijk is aan de wortel uit MS_{within} (in hoofdstuk 10 hebben we dit ontdekt). Dus als we de vinvergroting en controle als voorbeeld nemen, krijgen we

$$d = \frac{|\overline{Y}_1 - \overline{Y}_2|}{\sqrt{MS_{within}}} = \frac{|17{,}45 - 8{,}45|}{\sqrt{54{,}897}} = 1{,}21$$

Dit geldt als een heel groot effect. Bedenk even wat de uitkomst betekent. De standaardafwijking geeft aan hoe sterk individuele meerminnen verschillen. Hoe groot is het behandelingseffect in vergelijking hiermee? Word je veel tevredener over je lijf of schuif je in de populatie niet zoveel op? Volgens Cohens *d* bedraagt het verschil tussen de gemiddelden ruim 1,2 standaardafwijkingen.
Daarnaast hebben we nog *Hedges' g*. Volgens de meeste bronnen is deze schatter identiek aan Cohens *d*, maar dan met een correctie (Cohens *d* is namelijk een ietwat onzuivere schatter van δ). Deze correctie is nogal technisch en omslachtig om met de hand te verrichten, dus ik stel voor dat je hem alleen gebruikt als je een computer het werk kunt laten doen. Vanaf versie 27 kun je de hulp van SPSS inschakelen, mits je het algoritme voor de *ongepaarde t-toets* gebruikt (zie boek 1, hoofdstuk 9). Je hoeft daartoe enkel de nieuwe optie *Estimate effect sizes* aan te houden (deze staat standaard aangevinkt). Als we de vinvergroting met de controlegroep vergelijken, krijgen we de volgende extra tabel:

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
LICHAAMSBEELD	Cohen's d	8,636	1,042	,489	1,588
	Hedges' correction	8,754	1,028	,482	1,566
	Glass's delta	4,725	1,905	1,181	2,611

a. The denominator used in estimating the effect sizes.

Cohen's d uses the pooled standard deviation.

Hedges' correction uses the pooled standard deviation, plus a correction factor.

Glass's delta uses the sample standard deviation of the control group.

Ziedaar: Cohens *d*, de correctie van Hedges én Glass' delta (zometeen meer over die laatste). We krijgen zelfs voor elke schatter een betrouwbaarheidsinterval, hetgeen erg nuttig is. Jammer genoeg kleeft aan deze SPSS-uitvoer één nadeel wanneer je onderzoek uit meer dan 2 groepen bestaat: dan wordt s_p dus niet gebaseerd op alle groepen, alleen op de 2 groepen die je op dat moment vergelijkt. (Glass heeft hier geen last van; die gebruikt immers maar één standaardafwijking, zoals je hierna zult zien.)

Glass' $\hat{\Delta}$

Deze optie is beter als we *niet* willen uitgaan van gelijke varianties. Glass' $\hat{\Delta}$ (een hoofdletter delta, met een hoedje dat 'schatting' betekent) gebruikt niet de gepoolde standaardafwijking, maar die van één van de groepen:

$$\hat{\Delta} = \frac{|\overline{Y}_1 - \overline{Y}_2|}{s_1}$$

Die s_1 moet de standaardafwijking voorstellen van de 'basis'groep, zoals een controlegroep. Zo kunnen we zien hoe groot het behandelingseffect is, ten opzichte van individuele verschillen tussen onbehandelde deelnemers. In dat geval krijgen we:

$$\hat{\Delta} = \frac{|17{,}45 - 8{,}45|}{4{,}725} = 1{,}90$$

Het effect van een vinvergroting lijkt op basis hiervan nog groter – zeg maar gerust 'enorm'. Dit laat onverlet dat de standaardafwijking in de vinvergrotingsgroep heel groot was (11,262) vergeleken met die van de andere groepen. Anders gezegd: let erop dat al deze schatters van δ (met name Glass' $\hat{\Delta}$) alleen iets zeggen over de *gemiddelde* effectiviteit van een behandeling. Tussen *individuele* meerminnen kan het effect soms nog flink verschillen.

III Andere effectmaten

De tot nu toe besproken effectmaten zijn geschikt voor onderzoeken waarin we *groepen* (een *categorische* onafhankelijke variabele) vergelijken op een *kwantitatieve* afhankelijke variabele. Wat kunnen we het beste doen in andere designs?

- Is zowel je onafhankelijke als je afhankelijke variabele *kwantitatief*, dan ben je waarschijnlijk geïnteresseerd in hun *correlatie* (ρ) of het kwadraat hiervan, de *proportie verklaarde variantie* (ρ^2). Die laatste is equivalent aan η^2 in ANOVA-modellen.
  In boek 1 hebben we geleerd dat de proportie verklaarde varia(n)tie zich laat schatten met R^2. Eigenlijk doet zich hier precies dezelfde situatie voor. R^2 is hetzelfde als $\hat{\eta}^2$ en daarom onzuiver. Beter kun je de aangepaste R^2 gebruiken. De meest gebruikelijke formule is:

Aangepaste R^2

$$R^2_{aangepast} = 1 - \frac{MS_{residu}}{MS_{totaal}}$$

  De aangepaste R^2 gebruikt dus de varianties *(mean squares)* in plaats van de kwadratensommen, en hij kijkt min of meer naar de proportie variantie die *niet onverklaard* is. Met een beetje algebra valt echter te bewijzen dat dit hetzelfde is als:

$$R^2_{aangepast} = \frac{SS_{regressie} - df_{regressie} MS_{residu}}{SS_{totaal}}$$

  Met andere woorden: de aangepaste R^2 is identiek aan $\hat{\varepsilon}^2$. Nog een argument om bij ANOVA voor epsilon-kwadraat te gaan, lijkt me! ☺
- Zijn je onafhankelijke en afhankelijke variabele juist allebei *categorisch*? Dan vormt de *odds-ratio* een goede effectmaat (voor 2×2-tabellen). De odds-ratio in een aselecte steekproef is een zuivere schatter van de bijbehorende populatieparameter, dus dat schiet lekker op. Hierover meer in hoofdstuk 8.

Zijn er nog vragen?

$\hat{\eta}^2$, $\hat{\varepsilon}^2$ en $\hat{\omega}^2$ zijn allemaal steekproefschatters … maar we noteren ze stuk voor stuk met Griekse symbolen. Dat was de afspraak niet, Vincenzo!
Ja, sorry, het was niet mijn idee … Wat mij betreft moeten we Griekse symbolen reserveren voor parameters. Wat ook niet helpt is dat η (èta) en ω (omega) geen equivalent hebben in het Romeinse alfabet; de omega is bijvoorbeeld een lange 'o'. ε is al helemaal een ongelukkige keuze voor een letter, want die gebruikten we al voor de errorterm van ANOVA. ☺ Ik heb er maar hoedjes op gezet om te benadrukken dat het schatters zijn.

η^2 (de populatieparameter dus) is gebaseerd op varianties oftewel mean squares. Waarom nemen alle schatters als basis echter de kwadratensommen (sums of squares)?
Omdat hiervoor de ANOVA-vergelijking opgaat: $SS_{total} = SS_{between} + SS_{within}$. In een steekproef geldt het ANOVA-model niet voor de varianties (oftewel de *mean squares*). Kijk maar naar tabel 1.1: $MS_{total} \neq MS_{between} + MS_{within}$. (Deze ongelijkheid komt door de verschillende aantallen vrijheidsgraden van elke steekproefvariantie.) Vandaar ook het minuscule verschil in de naam: de steekproefschatter $\hat{\eta}^2$ noemt men doorgaans de proportie verklaarde variatie (zonder 'n'), maar de populatieparameter η^2 heet de proportie verklaarde variantie. Wat mij betreft mag je deze benamingen gerust door elkaar gebruiken. ☺

1.3 Zien is geloven? Assumpties opnieuw bekeken

Vereiste voorkennis	Handig om te hebben
Boek 1, hoofdstuk 9: Ongepaarde t-toets	Boek 1, hoofdstuk 10: Eenweg-ANOVA Boek 1, hoofdstuk 17: Enkelvoudige regressie

Bijna alle parametrische toetsen (denk aan t-toetsen, ANOVA en regressiemodellen) gaan uit van normaliteit, gelijke varianties en lineariteit. De tijd is rijp om hier grondiger naar te kijken, want sommige methoden om assumpties te controleren die we in boek 1 bespraken, waren nogal subjectief. Ben je klaar voor een diepe duik?

I We proberen gewoon normaal te doen

In hoofdstuk 10 van *Statistiek in stijl 1* bestudeerde onderzoeker Andersen $N = 87$ meerminnen die zich onzeker voelden over hun uiterlijk, en dit wilden compenseren met een vinvergroting. Een derde van hen kreeg de chirurgie waar ze om gevraagd hadden, een derde nam in plaats daarvan groepstherapie om een immaterialistische levenshouding te ontwikkelen, en de rest werd gedurende het experiment op een wachtlijst geplaatst. Twee maanden later mat Andersen met een vragenlijst hoe tevreden ze waren met hun lichaam.
De normaliteitsassumptie stelt dat elke steekproef getrokken is uit een normaal verdeelde populatie. Om eerlijk te zijn vormt die ideale wereld een zeldzaamheid. De meeste realistische populaties zullen *niet* exact normaal zijn. De echte vraag die we moeten stellen, is: maakt dat iets uit? Niet

Centrale-limietstelling

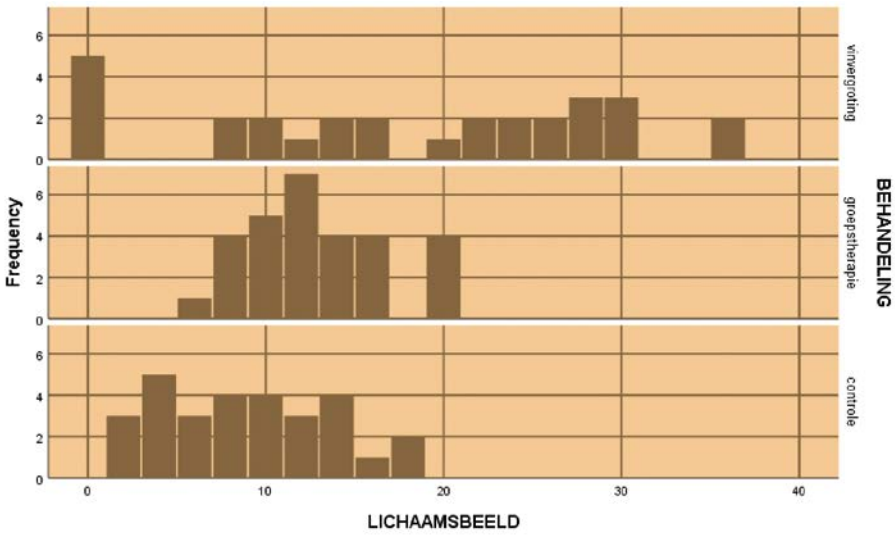
als de *steekproeven groot genoeg zijn*. Volgens de *centrale-limietstelling* wordt de *steekproevenverdeling* van een gemiddelde ongeveer normaal zolang je een grote steekproef trekt, zelfs als de populatieverdeling dat niet is. Denk hier altijd aan voordat je de moeite neemt de assumptie te beoordelen.

Dus wat is groot genoeg? In het algemeen geldt, hoe schever de populatie, des te groter moet je steekproef zijn om de steekproevenverdeling alsnog ongeveer normaal te maken. Hier heb je wat richtlijnen:

- Als $N \geq 15$, je geen uitschieters hebt, *en* je verdeling van steekproefscores er niet extreem scheef uitziet, mag je al leunen op de centrale-limietstelling. Beschouw dit als een soort ondergrens. Is je steekproef kleiner, vertrouw dan alleen extreme p-waarden (ver onder 0,01 of ver boven 0,10);
- In geval van sterke scheefheid zit je meestal veilig met $N = 25$. Een strengere richtlijn vraagt om $N = 40$;
- Vergelijk je meerdere groepen, dan is $n_i = 20$ per groep vaak genoeg, zelfs als zich een sterke scheefheid voordoet. $n_i = 15$ per groep is oké *mits* je aan de criteria uit het eerste punt voldoet.
- In *meervoudige-regressiemodellen* (zie hoofdstuk 10) moet je 10 of 15 proefpersonen verzamelen *per predictor*; het advies varieert een beetje van auteur tot auteur. Een model met 3 predictoren vereist bijvoorbeeld minstens $N = 30$ of $N = 45$ als de normaliteitsassumptie geschonden is. Merk op dat je voor genoeg *power* vaak sowieso een grotere steekproef nodig hebt dan dat.

In dit licht zien Andersens steekproeven er vrij groot uit, maar speciaal voor deze paragraaf gaan we normaliteit toch checken. De populaties hebben we niet, dus we kunnen alleen kijken naar de verdelingen van de tevredenheidsscores in de steekproef (figuur 1.3). Wat denk je, waarde lezer? Zien ze er acceptabel uit? Bij die laatste twee zou ik zeggen van wel, maar ik kan me voorstellen dat jij misschien niet zo zeker bent. Al met al is het meer nattevingerwerk dan je zou willen: 'Mmmjaaah … dit lijkt wel ongeveer op een normale verdeling … denk ik.' 😐

FIGUUR 1.3 Histogrammen van de steekproefscores



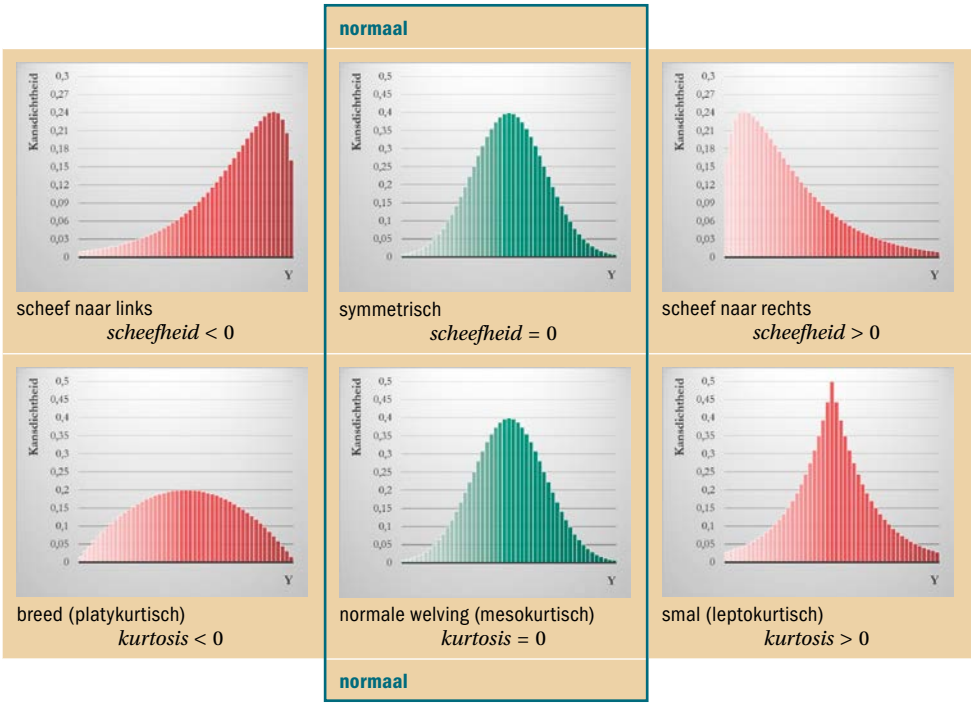
De tijd is dan ook gekomen om wat gevorderde middelen te onthullen voor een check op normaliteit. Ben je een student, vraag je docent dan welke methoden je moet leren. Ben je een onderzoeker, zoek dan hierna mijn adviezen op, want volgens mij zijn sommige methoden nuttiger dan andere. Ik raad aan om naar twee of drie dingen te kijken en te hopen dat ze hetzelfde zeggen. Heb je eenmaal je methoden gekozen, pas ze dan toe op *elke conditie in het onderzoek* (hier vinvergroting, groepstherapie en controle). Om het voorbeeld niet te langdradig te maken, zal ik hier alleen de eerste conditie uitwerken.

Scheefheid en kurtosis: vuistregel	
+ Supergemakkelijk te gebruiken. + We kunnen bepalen *hoe non-normaal* de populatie lijkt. Hoe verder de scheefheid of kurtosis zich buiten het gebied van [−1, +1] begeeft, des te groter moet de steekproef zijn om het probleem te verhelpen. – Deze vuistregel gebruikt alleen de steekproefschatting, die instabiel kan zijn als de steekproef klein is.	
Eindoordeel	**+**

Scheefheid betreft, verrassend genoeg, hoe scheef de verdeling van een variabele is. *Kurtosis* staat soms beschreven als de 'gepiektheid' van een verdeling. Dat laatste klopt niet helemaal – het gaat eigenlijk om de welving van de staarten – maar ach, voor nu volstaat het. Normale verdelingen hebben een scheefheid van 0 (ze zijn perfect symmetrisch) en ook een kurtosis van 0 (hun staarten zijn precies breedgewelfd genoeg). Figuur 1.4 laat zien wat ik bedoel.

Scheefheid
Kurtosis

FIGUUR 1.4 Scheefheid en kurtosis



De *Explore*-analyse in SPSS biedt een gespierde *Descriptives*-tabel die heel wat info bevat over je verdeling van steekproefscores. De scheefheid en kurtosis staan onderin vermeld.

1

Wil je meedoen? Ga dan naar *statistiekinstijl.noordhoff.nl*, log in bij de online leeromgeving en download de instructies en gegevensbestanden voor hoofdstuk 1, paragraaf 1.3.

Descriptives

	BEHANDELING			Statistic	Std. Error
LICHAAMSBEELD	vinvergroting	Mean		17,45	2,091
		95% Confidence Interval for Mean	Lower Bound	13,16	
			Upper Bound	21,73	
		5% Trimmed Mean		17,40	
		Median		20,00	
		Variance		126,828	
		Std. Deviation		11,262	
		Minimum		0	
		Maximum		36	
		Range		36	
		Interquartile Range		20	
		Skewness		-,228	,434
		Kurtosis		-1,109	,845

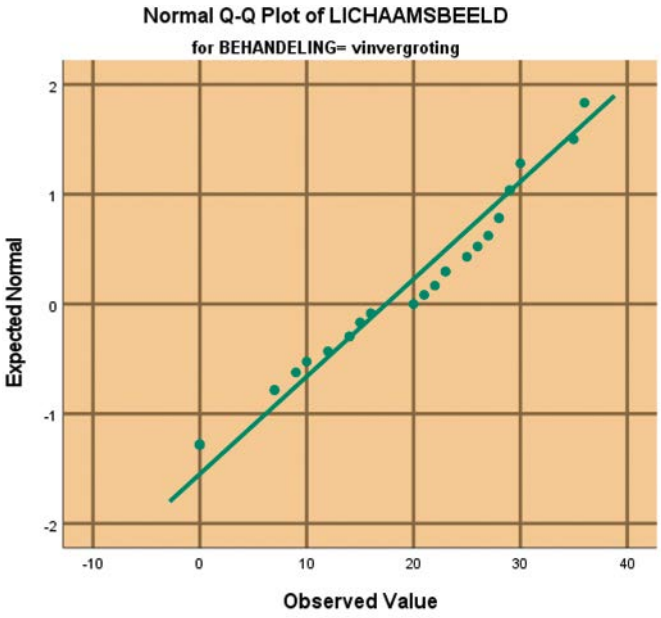
Het ziet ernaar uit dat het lichaamsbeeld van de 29 meerminnen in de vinvergrotingsgroep licht scheef naar links overhelde (–0,228), met vrij brede staarten (–1,109). Maar dit was slechts de steekproef; die kon een beetje afwijken van de populatie. Was het mogelijk dat de populatie toch een normale verdeling volgde (met een scheefheid en kurtosis van 0)? Onwaarschijnlijk, maar misschien bevond ze zich niet al te ver van dat ideaal … Een simpele vuistregel stelt dat *zowel scheefheid als kurtosis tussen –1 en +1 moet liggen*. Zolang dat het geval is, is de assumptie hooguit licht geschonden; steekproeven van beperkte omvang (zeg 15 proefpersonen) kunnen de situatie al rechtzetten. Als we deze regel volgen, lijkt het erop dat de populatie van de vinvergrotingsgroep nog vrij symmetrisch kon zijn, maar het is niet aannemelijk dat ze 'normaal gewelfd' was, want de kurtosis ligt nog lager dan –1. Gelukkig had Andersen 29 meerminnen, dus zijn ANOVA zou ook robuust moeten zijn tegen een iets slechtere kurtosis zoals deze.

Q-Q plot	
+ Nauwkeuriger dan dat je op het oog een histogram bekijkt (waarin kurtosis lastig te zien is). – Nog altijd vrij subjectief.	
Eindoordeel	**+/–**

In een normale verdeling zijn scores op een specifieke manier geordend. De meerderheid ligt in het centrum, en de frequentie van scores neemt vloeiend af naarmate ze zich verder van het gemiddelde bevinden. Een normaal *quantile-quantile plot* (kortweg Q-Q plot) vergelijkt jouw verdeling van steekproefscores met zo'n normale verdeling. Preciezer gezegd, het vergelijkt de posities die de steekproefscores hebben (op de x-as) met de posities die je ervan zou *verwachten*, als ze uit een normale verdeling afkomstig waren (op de y-as). De diagonale lijn door de grafiek (figuur 1.5) laat zien hoe goed de geobserveerde en de verwachte posities bij elkaar passen. De puntjes staan voor de individuele scores, en in dit geval volgen ze de diagonale lijn best goed. Hoe meer ze ervan afwijken, des te minder normaal ziet de verdeling eruit.

Quantile-quantile plot

FIGUUR 1.5 Q-Q plot



Normaliteitstoetsen (Kolmogorov-Smirnov/Shapiro-Wilk)	
+ Joepie, een toets! Nu kun je een conclusie trekken over de populatie! – … Maar dat wordt wel een alles-of-nietsconclusie. Normaliteit is of helemaal niet geschonden, of wel geschonden, maar dan kun je niet zeggen hoe ernstig. – Bedenkelijk bij kleine steekproeven (als het er echt om spant). De toets heeft misschien niet genoeg *power* om een schending van normaliteit te detecteren – dat wil zeggen, het kan gebeuren dat-ie niet significant is zelfs als de populatie *niet* normaal is. Gevaarlijk! – Bij grote steekproeven kan deze toets zoveel power hebben dat-ie al significant wordt bij een *net* niet normale populatie. Dan ga je je dus zorgen maken om een verwaarloosbaar probleem … vooral omdat de normaliteitsassumptie niets meer uitmaakt als je steekproef groot is.	
Eindoordeel	**+/–**

1

Misschien vraag je je af: kunnen we niet wat snelle *hypothesetoetsen* doen? Ja, dat kan. Er zijn een paar zogenoemde non-parametrische toetsen ontworpen door de *dream teams* Kolmogorov en Smirnov enerzijds, en Shapiro en Wilk anderzijds. Deze analysen gaan vooraf aan de ANOVA (of welk statistisch model je ook van plan was te gebruiken) en zijn specifiek bedoeld om de normaliteitsassumptie te beoordelen. In dit boek bespreek ik niet hoe ze werken, maar ik zal je wel vertellen hoe je de resultaten kunt interpreteren. Hun *nulhypothese* luidt (in woorden): *'De populatie is normaal verdeeld.'* Deze nulhypothese *mag niet verworpen worden*, anders is de assumptie aantoonbaar geschonden! SPSS trakteert ons op deze tabel in de *Explore*-analyse:

Tests of Normality

		Kolmogorov-Smirnov			Shapiro-Wilk		
	BEHANDELING	Statistic	df	Sig.	Statistic	df	Sig.
LICHAAMSBEELD	vinvergroting	,112	29	,200	,934	29	,068

Beide p-waarden (*Sig.* staat voor *Significance*) moeten we vergelijken met het standaard significantieniveau α, doorgaans 5% (0,05). Dan blijkt geen van beide toetsen significant (de Shapiro-Wilktoets komt echter in de buurt).

Andere methoden

Scheefheid en kurtosis: grove t-toetsen

Nog twee hypothesetoetsjes (parametrische ditmaal). Lees mijn algemene uitleg over scheefheid en kurtosis bij de vuistregel, mocht je dit nog niet gedaan hebben. De nulhypothese van elke toets is dat *de populatieparameter (scheefheid of kurtosis) gelijk is aan 0*. Als deze nulhypothese noch voor de scheefheid noch voor de kurtosis verworpen kan worden, kunnen we zeggen dat er aan de assumptie is voldaan (of op zijn minst dat ze niet duidelijk geschonden is). Om de toetsen te doen, regelen we als volgt twee t-waarden:

$$t = \frac{\textit{schatter (statistic)}}{\textit{standaardfout (standard error)}}$$

$$t_{scheefheid} = \frac{-0{,}228}{0{,}434} = -0{,}53 \qquad t_{kurtosis} = \frac{-1{,}109}{0{,}845} = -1{,}31$$

Je zou nu een precieze p-waarde kunnen opzoeken in de t-tabel op de website. Maar aangezien we dit grofweg doen, mag je zeggen dat *beide t-waarden tussen –2 en +2 moeten liggen*. Is er eentje extremer, dan is zijn toets significant (de p-waarde is kleiner dan 5%) en moet de nulhypothese worden verworpen. Dat wil je niet!

Waarom beveel ik dit niet aan? Noch scheefheid noch kurtosis volgt een normale steekproevenverdeling (ooit), dus de kritieke t-waarden zijn onjuist en kunnen leiden tot een verkeerde conclusie. (Daarbovenop lijden deze toetsen aan dezelfde zwaktes als hun broeders Kolmogorov-Smirnov en Shapiro-Wilk.)

Scheefheid en kurtosis: grove betrouwbaarheidsintervallen

De steekproefscheefheid was –0,228, maar waar zou de populatiescheefheid tussen liggen? Een *betrouwbaarheidsinterval* geeft antwoord op precies die vraag. We kunnen als volgt een grof exemplaar in elkaar zetten:

$$\textit{schatter (statistic)} \pm 2 \times \textit{standaardfout (standard error)}$$

Dit geldt voor zowel de scheefheid als de kurtosis, dus:

$$-0{,}228 \pm 2 \times 0{,}434 \rightarrow -0{,}228 \pm 0{,}868 \rightarrow [-1{,}096;\ 0{,}640]$$
$$-1{,}109 \pm 2 \times 0{,}845 \rightarrow -1{,}109 \pm 1{,}690 \rightarrow [-2{,}799;\ 0{,}581]$$

De scheefheid van de populatie lag dus waarschijnlijk tussen −1,096 en 0,640, en de populatiekurtosis viel waarschijnlijk tussen −2,799 en 0,581. *Als beide intervallen de waarde 0 bevatten*, kan de populatie de juiste kenmerken hebben gehad voor een normale verdeling. In dat geval (zoals hier) is er aan de assumptie voldaan (of op zijn minst is ze niet duidelijk geschonden). Merk op dat deze methode altijd leidt tot *dezelfde conclusie als de grove t-toetsen*.
Waarom beveel ik dit niet aan? Noch scheefheid noch kurtosis volgt een normale steekproevenverdeling (ooit), dus de grenzen van je interval zijn onjuist en kunnen leiden tot een verkeerde conclusie.

II We proberen gewoon gelijk te variëren

Volgende: de assumptie van gelijke varianties. Deze stelt dat alle steekproeven zijn getrokken uit populaties met dezelfde spreiding: de varianties zijn gelijk. Misschien herinner je je nog dat de *variantie* de tweede stap is in de berekening van een standaardafwijking: ze is σ in gekwadrateerde vorm. T-toetsen en ANOVA's nemen aan dat alle populaties dezelfde σ^2 hebben, want ze voegen de steekproefvarianties samen tot één schatting. Regressie gaat uit van iets heel vergelijkbaars (homoscedasticiteit; zie hoofdstuk 17 in boek 1 of hoofdstuk 10 voor meer informatie), maar de rest van deze subparagraaf gaat over de assumptie van t-toetsen en ANOVA's. Voordat we deze assumptie op de pijnbank leggen (hint: hij is meestal geschonden), moet je niet vergeten dat je je vaak de moeite kunt besparen. Als je steekproeven *ongeveer dezelfde omvang* hebben, is je toets robuust tegen een schending van gelijke varianties! Andersens groepen waren allemaal even groot (29 meerminnen), dus in dit geval mogen we de assumptie in feite negeren. Maar we gaan hem toch nakijken. En daarvoor bestaan een paar methoden (waarvan we de meeste al eens hebben gezien).

Steekproefstandaardafwijkingen: vuistregel	
+ Je kunt schatten *hoeveel* de populatiestandaarddeviaties verschillen (zolang je mijn versie van de regel gebruikt). Hoe meer de verhouding de 2 benadert of zelfs overschrijdt, des te meer moeten de steekproefgroottes op elkaar lijken om de toets robuust te maken. – Gebruikt alleen de steekproefschattingen, die instabiel kunnen zijn als de steekproef klein is.	
Eindoordeel	**+**

De vuistregel vergelijkt de steekproefstandaardafwijkingen. Als deze niet te veel verschillen, schatten ze mogelijk nog steeds dezelfde σ. De regel laat zich op meerdere manieren uitschrijven, maar ik ben enthousiast over een specifieke versie die uitdrukt *hoe ongelijk* de standaardafwijkingen zijn. Ze luidt $\frac{s_{max}}{s_{min}} < 2$ oftewel, *de grootste standaardafwijking gedeeld door de kleinste moet kleiner zijn dan 2*. In Andersens geval bleek (uit de data in boek 1) dat $\frac{11{,}262}{3{,}942} = 2{,}857 > 2$ … verdraaid. ☹ Dit houdt in dat de grootste standaardafwijking meer dan twee keer zo groot was als de kleinste – bijna

drie keer. Ze lagen zo ver uit elkaar dat ze waarschijnlijk *niet* dezelfde σ schatten. De vuistregel vertelt ons dus dat de assumptie geschonden is. We zagen het al aankomen, lijkt me.
Is de uitkomst van je berekening kleiner dan 2, dan zou de assumptie hooguit licht geschonden moeten zijn (er kan zelfs aan voldaan zijn, hoewel dat niet al te waarschijnlijk is). In dat geval mogen je steekproeven qua omvang redelijk van elkaar verschillen – maar hoe gelijker ze zijn, hoe beter.

Levene's toets	
+ Joepie, een toets! – Heeft dezelfde problemen als de toetsen voor de normaliteitsassumptie (zie eerder). – Is gevoelig voor normaliteit: als deze assumptie geschonden is, kan Levene een onjuiste uitslag geven (behalve als je steekproeven groot zijn).	
Eindoordeel	**+/–**

Ik zal je besparen hoe deze toets werkt in detail, wat hoogstwaarschijnlijk niet erg is, zolang je het resultaat kunt aflezen in SPSS-uitvoer:

Test of Homogeneity of Variances

LICHAAMSBEELD

Levene Statistic	df1	df2	Sig.
26,570	2	84	,000

Levene toetst de nulhypothese H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_i^2$: we hebben gelijke populatievarianties. Deze nulhypothese *mag niet verworpen worden*, anders is de assumptie geschonden! Check de p-waarde *(Sig.)* en hanteer een α of 0,05: als de toets significant is, wordt de nulhypothese verworpen en zijn de varianties aantoonbaar ongelijk. Dat is hier het geval: $p < 0{,}001$. ☹

Andere methode
F-toets op de varianties
Om te begrijpen wat hier gebeurt, moet je meer weten over ANOVA (zie hoofdstuk 10 in boek 1). Deze speciale versie van een F-toets deelt de twee steekproefvarianties door elkaar om een F-ratio te maken:

$$F = \frac{\textit{grootste variantie}}{\textit{kleinste variantie}} = \frac{11{,}262^2}{3{,}942^2} = 8{,}162$$

Hiermee toetsen we de nulhypothese

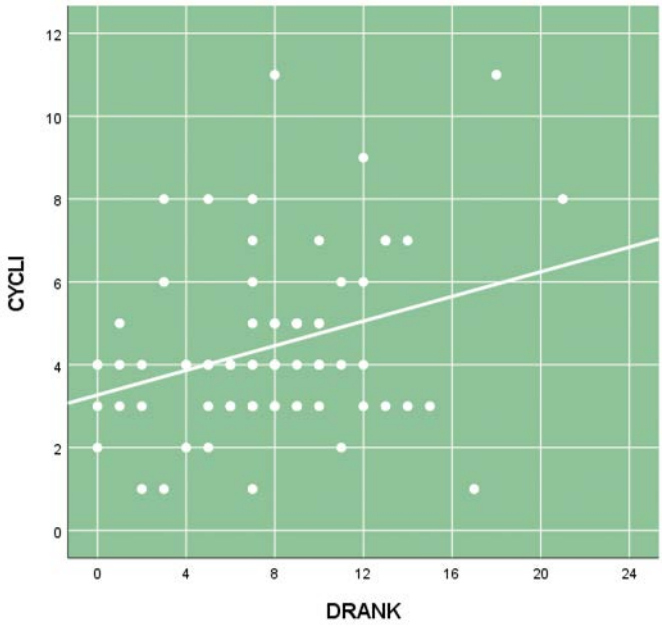
$$H_0: \sigma_1^2 = \sigma_2^2$$

Met een F-tabel kun je de p-waarde bepalen. De vrijheidsgraden bedragen $n_i - 1$ voor zowel de teller (bovenkant van de F-formule) als de noemer (onderkant), dus zoek de steekproefgrootten op (hier in beide gevallen 29). Dit zal doorgaans betekenen dat de F-tabel van *Statistiek in stijl* niet groot genoeg is (sorry!). Het blijkt hier dat $p < 0{,}001$, wat betekent dat het verschil tussen de steekproefvarianties significant is. We moeten de nulhypothese verwerpen … en dat wilden we niet, net zoals bij Levene's toets. Ook volgens deze laatste methode is de assumptie geschonden.
Waarom beveel ik dit niet aan? Deze toets lijdt aan dezelfde problemen als Levene's toets en is zelfs meer werk, want je kunt hem niet door SPSS laten uitvoeren.

III We proberen gewoon de zaak recht te zetten

In hoofdstuk 17 van boek 1 inspecteerde onderzoekster Rielgalad of alcoholconsumptie zorgt voor vruchtbaarheidsproblemen onder vrouwelijke Elfen. 74 aselecte Elfen die recent zwanger waren geworden, kregen de vraag hoeveel glazen lavendelmede zij wekelijks dronken, en hoeveel menstruatiecycli ze hadden doorlopen alvorens ze met succes bevrucht waren geraakt.
Laten we een blik werpen op het verband tussen drank en cycli. Bij lineaire regressie maken we een spreidingsdiagram (figuur 1.6) en we trekken er een rechte lijn doorheen om de algemene trend te beschrijven. Maar is zo'n lijn gepast? Of ziet de echte trend er eerder uit als een curve?

FIGUUR 1.6 Spreidingsdiagram van Rielgalads data



Om deze assumptie te controleren, kan ik twee methoden bedenken:

Spreidingsdiagram	
+ Makkelijk! – Kan subjectief zijn.	
Eindoordeel	**+/–**

Dit is de traditionele aanpak. Check het diagram en besluit of je er tevreden over bent. Mijn indruk van Rielgalads data is prima: volgens mij vat een rechte lijn de trend adequaat samen (wat inhoudt dat Elfen die meer drinken extra werk te doen hebben als ze zwanger willen worden).

Kwadratische toets	
+ Joepie, een toets! Nu kun je conclusies trekken over de populatie! + Onderzoekt het meest gebruikelijke non-lineaire verband. + Is veel eleganter dan de eerste methode (dus vertel erover als je indruk wilt maken op je date). – Heeft de gebruikelijke problemen van een statistische toets. Suggestie: kijk ook naar de gestandaardiseerde coëfficiënt van het kwadratische effect om te zien hoe sterk het lijkt. (Let op: voor een regressiemodel moet je de predictor dan wel eerst *centreren*. Zie hoofdstuk 10B.)	
Eindoordeel	**+**

Waarom *toetsen* we niet of het verband tussen X en Y kromlijnig is – bijvoorbeeld kwadratisch? Om te beginnen berekenen we de gekwadrateerde waarden van drank (dus X^2). In SPSS kun je hiervoor *Compute Variable* gebruiken. Vervolgens voeg je deze nieuwe variabele toe aan het regressiemodel.

Wil je meedoen? Ga dan naar *statistiekinstijl.noordhoff.nl*, log in bij de online leeromgeving en download de instructies en gegevensbestanden voor hoofdstuk 1.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	3,574	,650		5,497	,000	2,277	4,870
	DRANK	,051	,152	,107	,338	,736	-,251	,354
	DRANK_KWADRAAT	,006	,008	,218	,686	,495	-,011	,022

a. Dependent Variable: CYCLI

Zorg ervoor dat je zowel *DRANK* als *DRANK_KWADRAAT* toevoegt. Je kunt dan als volgt werken. Kijk eerst naar het hogereorde-effect (de kwadratische term). Is het significant, laat het dan in het model staan en concludeer dat er een kwadratisch verband bestaat. Is het niet significant, haal het dan eerst weg voordat je het lineaire effect van drank interpreteert. Uiteraard is drank in het kwadraat hier volkomen insignificant; het spreidingsdiagram suggereerde helemaal geen kromlijnigheid.
Mocht het kwadratische effect *wel* significant zijn, dan denk je misschien dat we een probleem hebben. Maar in feite … hebben we het zojuist opgelost. ☺ Met de kwadratische term erin hebben we nu het verband tussen X en Y correct gemodelleerd, dus we kunnen het zonder verdere problemen beschrijven. Toppie!

Zijn er nog vragen?

Ik heb ergens gelezen dat de kurtosis van een verdeling niet kleiner dan 0 kan worden. Hoe zit dat?
Eigenlijk heb ik het in dit hoofdstuk over *exces-kurtosis*. Die is gemakkelijker om mee te werken. De normale verdeling heeft een 'reguliere' kurtosis van 3, maar voor exces-kurtosis trekken we de waarde 3 ervan af om de grootheid rond 0 te centreren. SPSS rapporteert exces-kurtosis.

Om op normaliteit te controleren, beveel je de grove t-toetsen niet aan. Kun je ze toch een beetje uitleggen?
De complete formule is

$$t_{scheefheid} = \frac{steekproefscheefheid\ (schatter) - populatiescheefheid\ (H_0)}{standaardfout}$$. Maar volgens

de nulhypothese is de populatiescheefheid 0, dus deze term valt altijd weg. ☺ Idem dito voor de kurtosis!

O ja, nu snap ik hem. En die grove betrouwbaarheidsintervallen?
De precieze formule is *steekproefresultaat* (*statistic*) $\pm$ t^* $\times$ *standaardfout*, zoals gewoonlijk. We zetten de kritieke t-waarde enigszins voor de vuist weg op 2.

Stel dat de assumptie van gelijke varianties niet opgaat. Wanneer zijn de groepen dan qua omvang voldoende gelijk?
Exact gelijke groottes zijn het beste. Als de vuistregel duidt op een schending van gelijke varianties, mogen de steekproefgrootten nog steeds verschillen met ongeveer 10%. Hoe meer het verschil die grens overschrijdt, hoe onzuiverder de toets wordt.

De vuistregel voor gelijke varianties kunnen we op diverse manieren opschrijven, zeg je. Mogen we ook zeggen: '2 keer de kleinste s moet groter zijn dan de grootste'?
Ja hoor, zoals ik al aangaf in boek 1. Deze versie van de regel werkt even-goed prima, maar in tegenstelling tot die van mij kan ze geen *effectgrootte* uitdrukken en studenten klutsen de zin de hele tijd door elkaar.

Wil je toch iets meer vertellen over Levene's toets? Hoe werkt deze?
Eerst berekenen we van iedere proefpersoon het *absolute* verschil tussen diens eigen score en diens groepsgemiddelde. Als de populaties verschillen qua varianties, is dit absolute verschil *gemiddeld* groter in de ene populatie dan in de andere. We gebruiken het absolute verschil dus als een nieuwe variabele, en vergelijken hierop de groepen … met een ANOVA.
Op het eerste gezicht klinkt dit als een buitengewoon snugger idee, maar als je even doordenkt, zul je beseffen wat het betekent. Dit is waarom Levene's toets van normaliteit uitgaat, net zoals een reguliere ANOVA. En hij neemt aan dat de absolute verschillen gelijke varianties hebben … dus misschien moeten we een Levene's toets doen voor Levene's toets? Maar dan kunnen we voor de zekerheid maar beter een Levene's toets voor Levene's toets voor Levene's toets doen. En …

ANCOVA

Ik ben bezig met een ANCOVA (hoofdstuk 4) en heb subparagraaf 1.3.III gelezen om te controleren op lineariteit. Wat kan ik met mijn analyse het beste doen?
Bereken de gekwadrateerde waarden van de covariaat en voeg deze variabele als tweede covariaat toe aan het model. Significant? Laat de kwadratische term dan gewoon in het model zitten. Je hebt nu het probleem opgelost. ☺
Mocht je onzeker zijn of de kwadratische term belangrijk is (als-ie een tweeslachtige p-waarde heeft van bijvoorbeeld 0,08), dan kun je ook $\hat{\omega}^2$ of $\hat{\varepsilon}^2$ berekenen om zijn effectgrootte te zien.

1.4 Metamorfose: monotone transformaties

Vereiste voorkennis	Handig om te hebben
Boek 1, hoofdstuk 1: Gegevens in kaart	Boek 1, hoofdstuk 10: Eenweg-ANOVA Boek 1, hoofdstuk 17: Enkelvoudige regressie

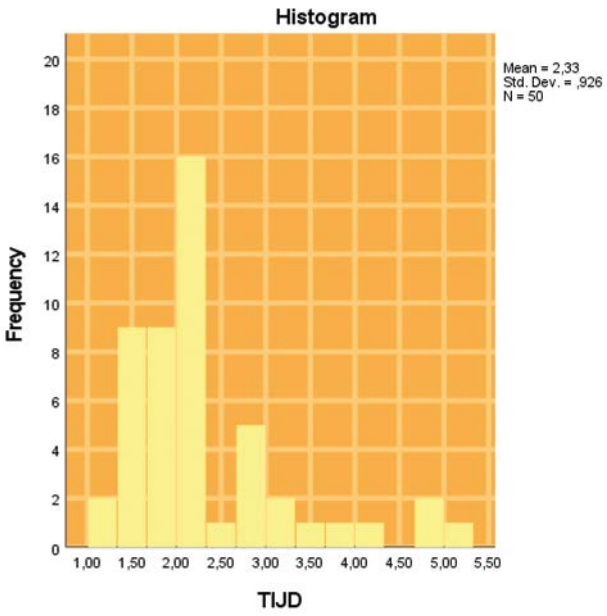
Lineaire *transformaties* zoals besproken in boek 1 veranderen niets aan de vorm van een verdeling, maar er bestaat nog een soort transformaties die dat wel doen. *Monotone* transformaties behouden de volgorde van de scores, maar scores mogen dichter bij elkaar komen of juist verder uit elkaar drijven. Met behulp hiervan kunnen we *scheve verdelingen symmetrischer maken*. Dit is handig wanneer je een statistische toets wenst uit te voeren, maar je tegen minstens één van de volgende problemen aanloopt:

- De assumptie van normaliteit lijkt geschonden en je steekproef is niet groot genoeg om de steun in te roepen van de centrale-limietstelling.
- De assumptie van gelijke varianties lijkt geschonden (t-toets of ANOVA) en je groepen zijn niet even groot.
- De assumptie van homoscedasticiteit lijkt geschonden (lineaire regressie).

De zaken gaan goed voor ijssalon Peach's Castle uit boek 1, dus recent heeft het bedrijf een bezorgdienst opgezet. Klanten binnen een straal van 5 kilometer kunnen nu hun ijs rechtstreeks aan huis laten leveren. Peach's Castle streeft er altijd naar zichzelf te verbeteren en trekt een aselecte steekproef van 50 bestellingen. De bezorger noteert de tijd die hij nodig heeft om bij iedere klant te arriveren. Deze tijd corrigeren we voor verschillen in afstand, door hem te delen door het aantal afgelegde kilometers; de variabele die hieruit komt, is de gebruikte tijd in minuten per kilometer. De resultaten volgen in een histogram (figuur 1.7).

Gemiddeld heeft de bezorger 2,33 minuten per kilometer nodig gehad. Dat is lekker snel; bewaard in een koelbox zal het ijs onderweg niet smelten. Echter, hij heeft ook een paar keer 5 minuten per kilometer verbruikt, en redelijk vaak tussen de 3 en 4 minuten. De verdeling is scheef naar rechts. Er bestaan vele monotone transformaties om de verdeling symmetrischer te maken. Ik bespreek drie populaire varianten. Deze zijn alle in staat om *rechtse scheefheid* te beperken, niet linkse scheefheid (lees hiervoor verder); ze hebben de neiging de ruimte tussen de scores te verkleinen, maar doen dit met meer enthousiasme bij de hogere scores.

FIGUUR 1.7 Verdeling van de tijdscores



Nog een belangrijke opmerking: wiskunde werkt zodanig dat deze datatransformaties het grootste effect sorteren als *de scores klein zijn (tussen 1 en 10) en de kleinste score gelijk is aan 1*. 'Echt waar?' In de bronvermelding van dit boek vind je een fraai artikel van de hand van Jason Osborne, waarin de nerds onder ons de wiskunde kunnen uitpluizen. Kortom, om aan deze voorwaarden te voldoen, kan het nodig zijn om eerst een lineaire transformatie toe te passen, voordat we een monotone doen. Laten we de tijdvariabele een beetje nader bestuderen:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TIJD	50	1,29	5,24	2,3334	,92551
Valid N (listwise)	50				

De huidige minimumwaarde is 1,29. Om dus de transformaties in deze paragraaf te optimaliseren, moeten we eerst alle scores met 0,29 opschuiven naar links. Hierdoor komt de minimumwaarde, ook bekend als het *ankerpunt*, op 1 terecht. Merk op dat de tijdscores vanaf dit punt moeilijker te interpreteren worden: ze geven niet langer aan hoeveel minuten de bezorger nodig had per kilometer, maar die hoeveelheid minuten minus 0,29 minuten. De aankomende transformaties gaan de theoretische betekenis achter de variabele verder vertroebelen. Dat is een nadeel dat we niet mogen negeren! Wél prettig is dat een hogere score nog steeds zal staan voor meer tijd (daarom spreken we van monotone transformaties).

Ankerpunt

Trekken we 0,29 af van alle scores (in SPSS kunnen we dat doen met *Compute Variable*), dan krijgen we dit:

Wil je meedoen? Ga dan naar *statistiekinstijl.noordhoff.nl*, log in bij de online leeromgeving en download de instructies en gegevensbestanden voor hoofdstuk 1, paragraaf 1.4.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TIJD_anker	50	1,00	4,95	2,0434	,92551
Valid N (listwise)	50				

Yep, dat is volgens plan gegaan. Je kunt zien dat de scores nu van 1,00 tot 4,95 reiken, dus we zijn klaar voor de start! Als je een variabele hebt waarvan het bereik ver over de 10 gaat, kun je de scores ook nog delen door een bepaald getal voordat je ze verankert op 1. Een paar voorbeelden:

- Ligt je bereik rond 50: deel de scores door 5 en veranker ze daarna. (Dit werkt niet perfect, want bij het verankeren schuiven de scores ook weer wat op, maar het zou voldoende moeten zijn.)
- Ligt je bereik rond 100: deel de scores eerst door 10.

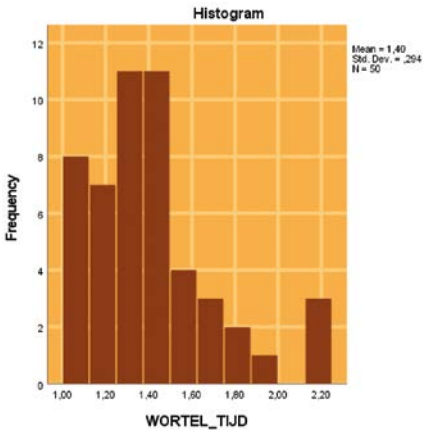
Dit hoeft hier niet. Aan de slag dan maar.

I Worteltransformatie

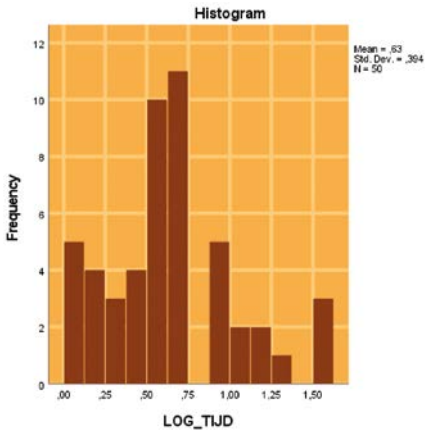
Deze spreekt voor zich: we trekken de *wortel* uit alle waarden. Om dit te doen, moet er voldaan zijn aan een paar voorwaarden die we al hebben afgedekt (zie *Zijn er nog vragen?*), door de tijden te verankeren op 1. Hier komt dus het histogram dat resulteert uit de worteltrekking (figuur 1.8).

Dit ziet er redelijk effectief uit; de verdeling is inderdaad een beetje symmetrischer geworden. Er is echter ruimte voor verbetering. Misschien werkt een andere transformatie beter?

FIGUUR 1.8 Worteltransformatie



FIGUUR 1.9 Logaritmische transformatie



II Logaritmische transformatie

De idee van een logaritme bespreek ik uitgebreid in hoofdstuk 8, als een voorbereiding op logistische regressie in hoofdstuk 11. Hier een korte herhaling: als $3^4 = 81$ (het is maar een voorbeeld), kunnen we ons afvragen tot welke macht we 3 moeten verheffen om 81 te krijgen. Dit is de logaritme van 81, met grondtal 3: $\log_3 81 = 4$. Als alternatief kunnen we het speciale wiskundige getal e als grondtal gebruiken ($e \approx 2{,}72$). De zogenoemde natuurlijke logaritme $\log_e X$ schrijven we ook wel op als $\ln X$. En vandaag is X de tijdvariabele.
Laten we de natuurlijke logaritme nemen van de vers verankerde tijdvariabele (gebruik deze en je voldoet aan dezelfde vereisten als voor een wortel-transformatie). Die ziet er goed uit (figuur 1.9)! Zoals je ziet heeft een log-transformatie wat meer pit. Een potentieel nadeel hiervan is dat-ie ook te ver kan gaan, en een verdeling kan produceren die scheef naar links is.

III Inverse transformatie

Het derde en krachtigste alternatief is dat we de *inverse* berekenen van de tijdvariabele. Voorbeeldje: de inverse van 2 is $\frac{1}{2}$ oftewel 0,5.

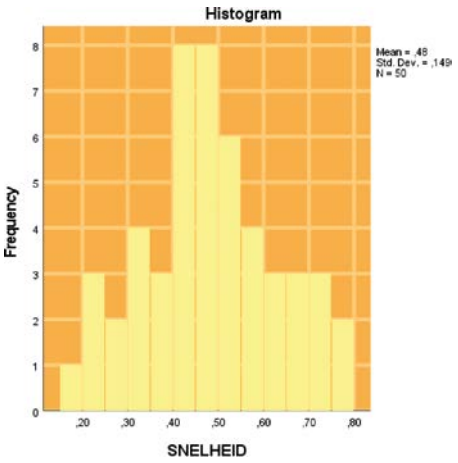
Voordat we verdergaan, wil ik je graag ergens op wijzen. De tijdvariabele vertegenwoordigt hoeveel minuten de bezorger erover deed om één kilometer af te leggen. Stel dat we hiervan de rechtstreekse inverse pakken, zonder iets anders te veranderen … wat krijgen we dan?

$$\frac{min}{km} \rightarrow \frac{1}{\left(\frac{min}{km}\right)} = \frac{km}{min}$$

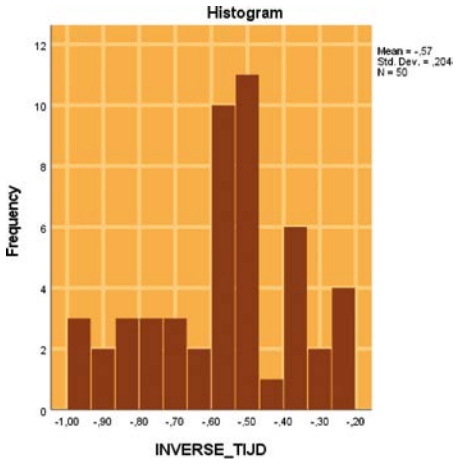
We krijgen een maat voor zijn *snelheid* (figuur 1.10)! En inderdaad: als de tijd die hij voor een bepaalde bestelling nodig heeft gelijk is aan 5 minuten per kilometer, bedraagt zijn snelheid 0,2 kilometer per minuut.

Hopelijk bewijst dit voor je, waarde lezer, dat monotone transformaties niets wezenlijks veranderen aan de oorspronkelijke variabele. Ze drukken de informatie die deze variabele bevat gewoon anders uit. We doen deze transformaties enkel om de verdelingsvorm van de scores te veranderen, zodat statistische toetsen die we op deze variabele verrichten geldig zullen zijn. Als de tijd van de bezorger niet normaal verdeeld is, geldt dat misschien wel voor zijn snelheid. In dat geval toetsen we gewoon hoe zijn snelheid beïnvloed wordt door bepaalde onafhankelijke variabelen! Daardoor zal het statistische model (zoals ANOVA of regressie) de juiste p-waarden en betrouwbaarheidsintervallen produceren.

FIGUUR 1.10 De snelheidsvariabele



FIGUUR 1.11 Inverse transformatie (verankerd op 1)



De situatie die ik zojuist beschreef is een aangename uitzondering, waarin een monotone transformatie de variabele theoretisch perfect interpreteerbaar houdt. Meestal is dit niet het geval, dus laten we dieper ingaan op de 'gebruikelijke' inverse. We kunnen deze benaderen zoals de andere methoden: eerst verankeren we de variabele op 1 en passen we de transformatie toe. Maar er is sprake van een extraatje. Pak je de inverse, dan draait de volgorde van de scores zich om: de hoogste wordt de laagste en vice versa. (Tenslotte is de langste tijd gekoppeld aan de laagste snelheid.) Om dit ongedaan te maken, spiegel je de schaal gewoon nog eens: *zet aan het begin van de berekening een minteken*.
Gelukt (figuur 1.11)! Deze verdeling is lichtjes scheef naar links en daarmee iets te enthousiast, maar ik zou hem geslaagd noemen.

IV Zo verhelp je linkse scheefheid

Is jouw verdeling niet scheef naar rechts maar scheef naar links, *vermenigvuldig ze dan gewoon met –1* voordat je een van de voorgaande transformaties verricht. Let erop dat je hiermee de volgorde van de scores onderste-boven kiepert, dus om hun interpretatie te versimpelen (die sowieso lastiger is), kun je ze na afloop van de transformatie nogmaals met –1 vermenigvuldigen.

V Resultaten

Op basis van onze verkenningen kunnen we een van de voorgaande alternatieven kiezen. In dit geval doet zowel de logaritmische als de inverse transformatie het goed, dus een van die variabelen laat zich goed gebruiken in onze statistische modellen. Deze conclusie vindt steun bij aanvullende statistieken over de histogrammen, zoals scheefheid en kurtosis (gebruik de gereedschapskist uit paragraaf 1.3). Bij wijze van uitzondering zal ik geen van de drie selecteren en in plaats daarvan voor de snelheidsvariabele gaan: een transformatie die zich zo natuurlijk laat interpreteren is normaal niet beschikbaar, maar nu wel, dus we kunnen er maar beter van gebruikmaken. En laten we wel wezen: zijn verdeling ziet er prima uit! ☺

TABEL 1.2 Scheefheid en kurtosis van de getransformeerde variabelen

	Wortel_tijd	Log_tijd	Inverse_tijd	Snelheid
Scheefheid	1,143	0,646	−0,270	−0,047
Kurtosis	1,146	0,165	−0,430	−0,462

Peach's Castle wil graag weten of het ijs meer vertraging oploopt in de spitsuren. Vandaar dat men bijhoudt welke bestellingen afgeleverd worden in het midden van de avondspits, en welke daarvoor of daarna. Er verrijzen twee condities.

Descriptives

TIJD

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
buiten de spits	32	2,1284	,56880	,10055	1,9234	2,3335	1,33	3,71
in de spits	18	2,6978	1,28815	,30362	2,0572	3,3384	1,29	5,24
Total	50	2,3334	,92551	,13089	2,0704	2,5964	1,29	5,24

ANOVA

TIJD

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,734	1	3,734	4,688	,035
Within Groups	38,238	48	,797		
Total	41,972	49			

ANOVA op de oorspronkelijke tijdvariabele levert een significant verschil op tussen de twee condities ($p = 0{,}035$); de bezorger lijkt in de spits meer tijd nodig te hebben. Is deze toets valide? Hoewel normaliteit misschien niet zo'n groot probleem vormt (de tweede steekproef is niet bijster groot, maar lichte scheefheid doet geen pijn), hebben de groepen duidelijk *ongelijke varianties* volgens de vuistregel. (In de steekproef was $\frac{s_{max}}{s_{min}} > 2$. Daarnaast was Levene's toets, hier niet weergegeven, zeer significant ($p < 0{,}001$).) Nu de ANOVA groepen vergelijkt die sterk van omvang verschillen, is-ie *niet* robuust tegen deze schending. Het verschijnsel dat we hier tegenkomen, doet zich vaker voor: scheefheid leidt wel eens tot ongelijke varianties, omdat het meestal een specifiek niveau van X is (hier de verkeerssituatie) dat voor de hogere scores zorgt.

Zoals altijd leidt een regressieanalyse tot een identieke uitslag (ik heb de uitvoer beperkt tot de *Coefficients*-tabel):

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	2,128	,158		13,490	,000	1,811	2,446
	VERKEER	,569	,263	,298	2,165	,035	,041	1,098

a. Dependent Variable: TIJD

Het betrouwbaarheidsinterval is interessant: het vertelt ons tussen welke waarden het tijdsverschil kan liggen in de populatie. Dit interval is echter gebaseerd op een normale steekproevenverdeling met een gepoolde standaardfout, net zoals de t-toets voor de helling en de ANOVA. De waarden die eruitkomen, zijn daarom niet te vertrouwen.
Laten we het nu nog eens proberen met als afhankelijke variabele de snelheid, het invers getransformeerde equivalent van de tijd:

Descriptives

SNELHEID

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
buiten de spits	32	,5013	,12807	,02264	,4552	,5475	,27	,75
in de spits	18	,4460	,17930	,04226	,3569	,5352	,19	,78
Total	50	,4814	,14916	,02109	,4390	,5238	,19	,78

ANOVA

SNELHEID

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,035	1	,035	1,603	,212
Within Groups	1,055	48	,022		
Total	1,090	49			

In deze analyse is het effect van de verkeerssituatie niet langer significant (p = 0,212); Peach's Castle heeft bij nader inzien geen bewijs dat de bezorger tijdens spitsuren langzamer is. De varianties zien er niet langer ongelijk uit (nu blijft de vuistregel overeind) en we weten dat de snelheidsvariabele mooi symmetrisch is, dus wat mij betreft kunnen we erop vertrouwen dat deze ANOVA ons een veel nauwkeuriger beeld geeft van de populatie.
Regressieanalyse zegt hetzelfde:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	,501	,026		19,130	,000	,449	,554
	VERKEER	-,055	,044	-,180	-1,266	,212	-,143	,033

a. Dependent Variable: SNELHEID

Volgens het betrouwbaarheidsinterval ligt het gemiddelde snelheidsverschil op populatieniveau waarschijnlijk tussen –0,143 en 0,033 kilometer per minuut. Dit interval is zuiverder dan zijn voorganger.
Het is ook mogelijk om de modellen te vergelijken op basis van gestandaardiseerde maten voor de effectgrootte (zie paragraaf 1.2). In dit geval ziet het verkeerseffect er volgens al die maten kleiner uit in het model met de getransformeerde variabele.

TABEL 1.3 Effectmaten volgens de twee verschillende modellen

	Cohens *d*	$\hat{\varepsilon}^2$	Gestandaardiseerde *b*
Tijd-model	0,64	0,070	0,298
Snelheid-model	0,37	0,012	–0,180

Houd in het achterhoofd dat een monotone datatransformatie in de meeste onderzoeken zorgt voor abstracte schattingen en betrouwbaarheidsintervallen. Dit kan een duidelijk nadeel zijn, want het is behoorlijk ingewikkeld om de schattingen terug te transformeren. Nu weet je tenminste waar je aan begint.

Zijn er nog vragen?

Een transformatie is dus monotoon als de hoogste score de hoogste blijft en de laagste score de laagste?
Om precies te zijn, gelden transformaties ook als monotoon als de laagste score de hoogste wordt. Waar het hier om gaat, is dat de volgorde van de scores hetzelfde blijft. Dat zou niet het geval zijn als na de transformatie bijvoorbeeld een score die eerst in het midden lag, opeens de laagste werd.

Je hebt aangegeven dat we aan een paar voorwaarden moeten voldoen om een wortel-transformatie te kunnen uitvoeren. Het verankeren van de variabele op 1 zorgt hiervoor. Wat zijn die voorwaarden?
Ten eerste moeten scores *groter dan 0* zijn; de wortel uit negatieve reële getallen bestaat tenslotte niet. Ten tweede mogen er *geen scores tussen 0 en 1* zijn; anders werkt de transformatie niet naar behoren. De wortel uit getallen tussen 0 en 1 zal groter zijn dan de oorspronkelijke getallen in plaats van kleiner, waardoor een deel van de schaal zich omkeert.

Ik snap het. En dit geldt dus ook voor een logaritmische transformatie?
Ja: de logaritme van negatieve getallen (en van 0) bestaat niet en op getallen tussen 0 en 1 is het effect van de transformatie buitensporig groot, vergeleken met getallen van 1 of hoger. Je kunt er dan beter voor zorgen dat alle getallen boven de 1 zitten.

Samenvatting

I Effectgrootten schatten

Welke effectparameter wil ik weten?

		Onafhankelijke variabele (*X*)	
		categorisch	**kwantitatief**
Afhankelijke variabele (*Y*)	**categorisch**	ζ (odds-ratio; zie hoofdstuk 8)	ζ (odds-ratio; zie hoofdstuk 8)
	kwantitatief	η^2 (proportie verklaarde variantie) of δ (gestandaardiseerd gemiddeld verschil)	ρ (correlatie) of ρ^2 (proportie verklaarde variantie)

Wat is de beste schatter?

Parameter	Schatters	Advies
Proportie verklaarde variantie (ANOVA) η^2	$\hat{\eta}^2$ $\hat{\varepsilon}^2$ $\hat{\omega}^2$	Gebruik niet $\hat{\eta}^2$ (die overschat η^2 systematisch). $\hat{\varepsilon}^2$ is de beste optie (die onderschat η^2 vaak enigszins). Ook $\hat{\omega}^2$ is een prima keuze (die onderschat η^2 doorgaans nog iets meer).
Gestandaardiseerd gemiddeld verschil (ongepaarde t-toets) δ	Cohens *d* Hedges' *g* Glass' $\hat{\Delta}$	Gebruik *d* of (nog iets beter) *g* als je durft uit te gaan van gelijke varianties. Doe je dat liever niet, ga dan voor Glass' $\hat{\Delta}$.
Proportie verklaarde variantie (regressie) ρ^2	R^2 Aangepaste R^2	Gebruik niet de reguliere R^2 (die overschat ρ^2 systematisch). De aangepaste R^2 vormt de zuiverste schatter.

II Monotone transformaties

Naam	Berekening	Relatieve kracht
Worteltransformatie	$X_i' = \sqrt{X_i}$	Mild
Logaritmische transformatie	$X_i' = \ln X_i$	Gematigd
Inverse transformatie	$X_i' = -1/X_i$	Agressief

In de genoemde formules staat *i* voor de proefpersoon.
Deze transformaties reduceren *rechtse scheefheid*. Wil je linkse scheefheid reduceren, vermenigvuldig dan voor *en* na de transformatie de scores met –1.
Zorg er vóór de transformatie voor dat alle scores *tussen 1 en 10* liggen (ongeveer) en zet het *ankerpunt op 1*.

Hoe nu verder?

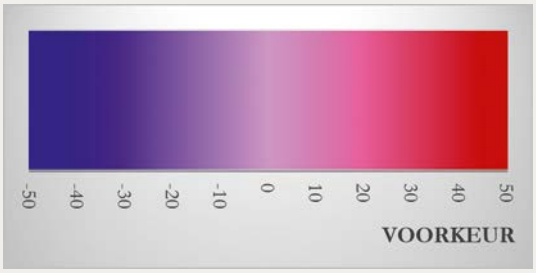
Cursus voor gorilla's
Hoofdstuk 2: Vergelijkingen en contrasten

1A Opdrachten

Hoe gemakkelijk krijg je in de sport een kind aan je zijde? Om dit uit te zoeken, nodigden Koeman et al. (1997) een aselecte steekproef van 93 kinderen tussen de acht en tien jaar uit voor een experiment. Zij stelden hen voor aan twee rivaliserende voetbalteams: de Razende Leeuwen en de Adelaars van Staal. Een videodocumentaire van vijf minuten vertelde de proefpersoontjes over de epische strijd die de teams al tientallen jaren voerden om de beker. Daarna werden de kinderen in drie groepen verdeeld. Koters in de 'rode' groep maakten een busreisje naar het stadion van de Razende Leeuwen, waar ze de spelers ontmoetten en naar een wedstrijd mochten kijken tussen de Razende Leeuwen en een anoniem team (door matchfixing wonnen de Razende Leeuwen met vlag en wimpel). Kinderen in de 'blauwe' groep daarentegen bezochten het stadion van de Adelaars van Staal, waar de Adelaars op precies dezelfde wijze een heldenrol op zich namen. Een derde groep ('neutraal') ging naar weer een ander stadion waar twee anonieme teams (niet verbonden aan de Adelaars of de Leeuwen) een vriendschappelijke wedstrijd speelden.
Na afloop kregen alle kinderen een voetbalshirt van het team dat ze hadden bezocht. De volgende dag moesten ze van Koeman een korte vragenlijst invullen. Met hun antwoorden bepaalde de onderzoeker de mate waarin ieder kind een voorkeur had voor één van de teams. Deze voorkeur

drukte hij uit op een schaal van –50 tot +50. Hierbij duidde –50 op een blinde loyaliteit aan de Adelaars van Staal, +50 op een aanbidding van de Razende Leeuwen, en 0 op een neutraal standpunt (figuur 1.12).

FIGUUR 1.12 De voorkeursschaal



Opdracht 1

In het eerste deel van Bijlage 1 vind je een dataverkenning van de voorkeursscores in de blauwe groep. Volgens welke criteria bevat deze groep uitschieters? Schop *alle* goede antwoorden het doel in. Sommige criteria zijn een beetje subjectief, dus leg in dat geval je beoordeling uit.

- ⚽ Eilandjes in het histogram
- ⚽ De lijst met extreme waarden
- ⚽ Het IQR-criterium
- ⚽ Het 5% getrimde gemiddelde
- ⚽ Z-scores

Opdracht 2

Voor de zekerheid vroegen Koeman et al. de kinderen aan het eind van de vragenlijst: *Vond je alle vragen duidelijk?* Proefpersoon nummer 54 antwoordde hierop 'nee'. Mochten de onderzoekers deze deelnemer verwijderen?

a Ja, want het was een uitschieter (een extreme fan van de Razende Leeuwen)
b Ja, dat zou zelfs verstandig zijn als dit kind geen uitschieter was
c Nee, trimmen of winsoriseren vormt altijd een betere oplossing

Opdracht 3

In het tweede deel van de Bijlage zijn de data wat opgeschoond. De *Descriptives*-tabel geeft informatie over alle drie de groepen. Koeman wilde hun gemiddelden vergelijken met een eenweg-ANOVA. Lijkt er sprake van een schending van gelijke varianties?

a Ja, dus er is eerst een datatransformatie nodig
b Ja, maar de ANOVA was vrij robuust tegen deze schending
c Nee, want de groepen zijn ongeveer even groot

Opdracht 4

Was de normaliteitsassumptie voor deze data geschonden? Gebruik de criteria die jouw docent aanbeveelt of volg de adviezen in dit hoofdstuk op. Zo ja, vormde deze schending een probleem voor de analyse?

Opdracht 5

Waarom is een datatransformatie van de voorkeursscores in dit onderzoek geen vruchtbare optie?

a Deze variabele kan negatieve waarden aannemen
b De vierkantswortel, logaritme of inverse van een voorkeursscore heeft geen praktische betekenis
c De verdeling van deze variabele is in de blauwe groep scheef naar rechts, maar in de rode groep juist scheef naar links en in de neutrale groep vrij symmetrisch

Opdracht 6

6.1 Bestudeer de ANOVA in Bijlage 1. Volgens de richtlijnen van Cohen is het algemene groepseffect naar schatting …

a Klein
b Medium
c Groot

6.2 Waarom zegt η^2 in dit onderzoek nog niet alles over het groepseffect?

6.3 Bereken Cohens d voor elk groepspaar (blauw versus neutraal, rood versus neutraal en blauw versus rood). Welke twee groepen verschillen het meest van elkaar?

6.4 Welk bezwaar zou Koeman in dit geval tegen Cohens d kunnen maken? (Hint: deze effectschatter rust op een bepaalde assumptie.)

6.5 Dan maar Glass' $\hat{\Delta}$. Bereken ook deze schatter voor elk groepspaar. Tegen welk dilemma loop je aan?

De uitwerkingen vind je op de website, statistiekinstijl.noordhoff.nl.

Bijlage 1 (Koeman)

Dataverkenning blauwe groep

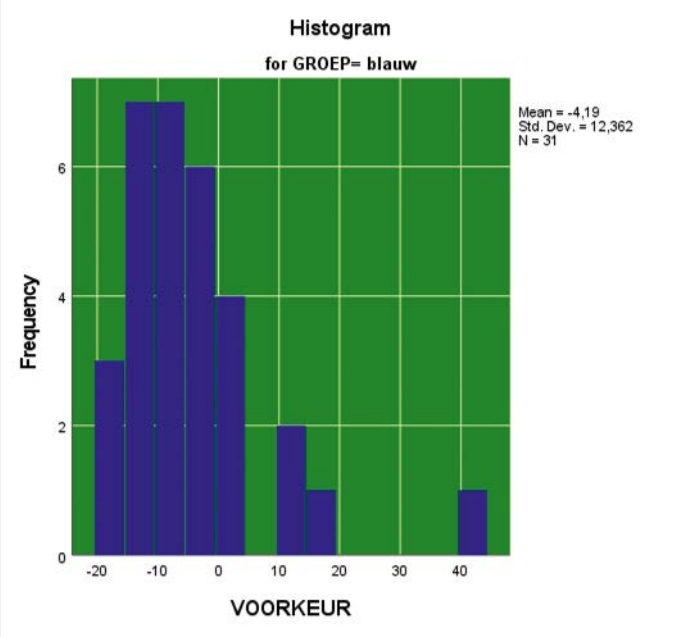
Descriptives

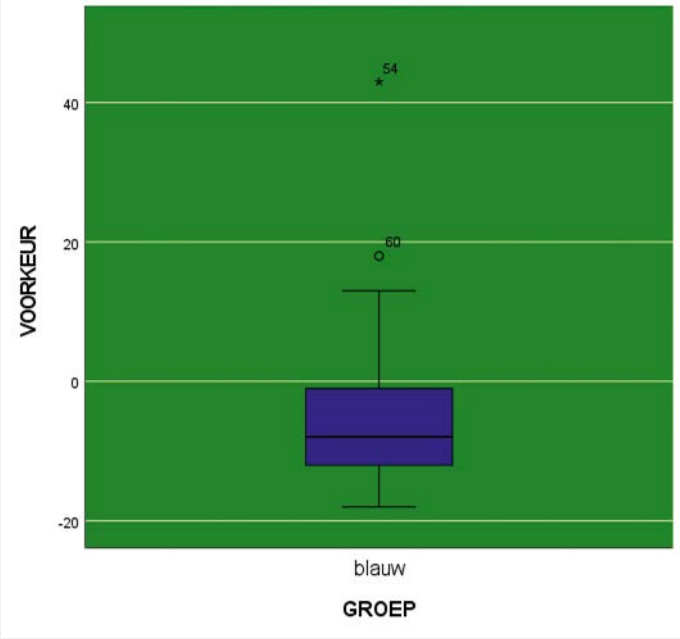
	GROEP			Statistic	Std. Error
VOORKEUR	blauw	Mean		-4,19	2,220
		95% Confidence Interval for Mean	Lower Bound	-8,73	
			Upper Bound	,34	
		5% Trimmed Mean		-5,58	
		Median		-8,00	
		Variance		152,828	
		Std. Deviation		12,362	
		Minimum		-18	
		Maximum		43	
		Range		61	
		Interquartile Range		12	
		Skewness		2,140	,421
		Kurtosis		6,219	,821

Extreme Values

	GROEP			Case Number	Value
VOORKEUR	blauw	Highest	1	54	43
			2	60	18
			3	33	13
			4	32	10
			5	46	4[a]
		Lowest	1	45	-18
			2	53	-17
			3	43	-16
			4	51	-15
			5	41	-15

a. Only a partial list of cases with the value 4 are shown in the table of upper extremes.


Histogram
for GROEP= blauw
Mean = -4,19
Std. Dev. = 12,362
N = 31
Frequency
0
2
4
6
-20
-10
0
10
20
30
40
VOORKEUR


54
60
VOORKEUR
-20
0
20
40
blauw
GROEP

ZVOORKEUR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1,14226	1	3,2	3,2	3,2
	-1,08138	1	3,2	3,2	6,5
	-1,02050	1	3,2	3,2	9,7
	-,95963	2	6,5	6,5	16,1
	-,83787	1	3,2	3,2	19,4
	-,77700	3	9,7	9,7	29,0
	-,71612	1	3,2	3,2	32,3
	-,65524	1	3,2	3,2	35,5
	-,59437	2	6,5	6,5	41,9
	-,53349	3	9,7	9,7	51,6
	-,47261	1	3,2	3,2	54,8
	-,35086	4	12,9	12,9	67,7
	-,16823	2	6,5	6,5	74,2
	-,04648	1	3,2	3,2	77,4
	,07528	1	3,2	3,2	80,6
	,19703	2	6,5	6,5	87,1
	,56229	1	3,2	3,2	90,3
	,74492	1	3,2	3,2	93,5
	1,04930	1	3,2	3,2	96,8
	2,57122	1	3,2	3,2	100,0
	Total	31	100,0	100,0	

Verkenning opgeschoonde data

Descriptives

	GROEP			Statistic	Std. Error
VOORKEUR	blauw	Mean		-5,77	1,620
		95% Confidence Interval for Mean	Lower Bound	-9,08	
			Upper Bound	-2,45	
		5% Trimmed Mean		-6,33	
		Median		-8,00	
		Variance		78,737	
		Std. Deviation		8,873	
		Interquartile Range		11	
		Skewness		1,019	,427
		Kurtosis		,773	,833
	neutraal	Mean		-,87	4,049
		95% Confidence Interval for Mean	Lower Bound	-9,14	
			Upper Bound	7,40	
		5% Trimmed Mean		-1,47	
		Median		-2,00	
		Variance		508,316	
		Std. Deviation		22,546	
		Interquartile Range		36	
		Skewness		,291	,421
		Kurtosis		-,565	,821
	rood	Mean		8,63	1,212
		95% Confidence Interval for Mean	Lower Bound	6,15	
			Upper Bound	11,11	
		5% Trimmed Mean		9,06	
		Median		9,00	
		Variance		44,102	
		Std. Deviation		6,641	
		Interquartile Range		8	
		Skewness		-,920	,427
		Kurtosis		,555	,833

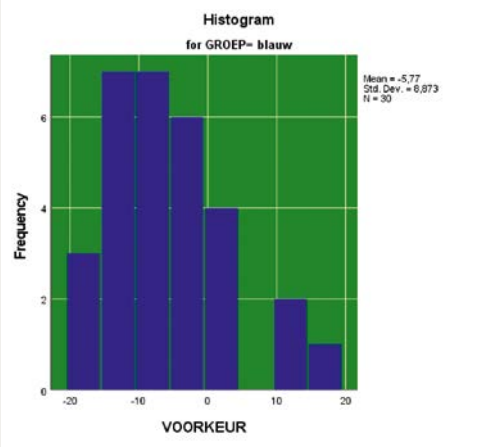
1

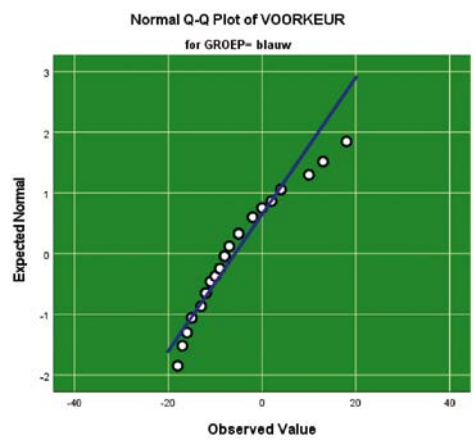
Tests of Normality

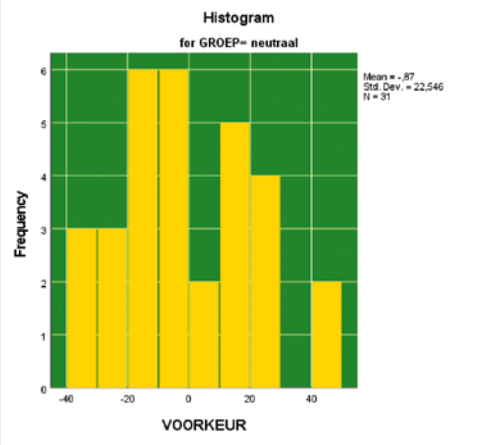
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	GROEP	Statistic	df	Sig.	Statistic	df	Sig.
VOORKEUR	blauw	,166	30	,035	,924	30	,035
	neutraal	,083	31	,200*	,969	31	,479
	rood	,146	30	,103	,918	30	,024

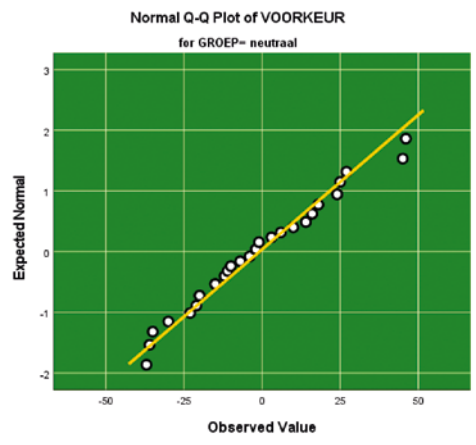
*. This is a lower bound of the true significance.

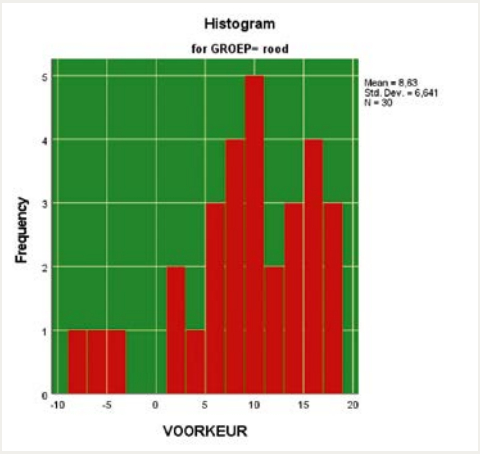
a. Lilliefors Significance Correction

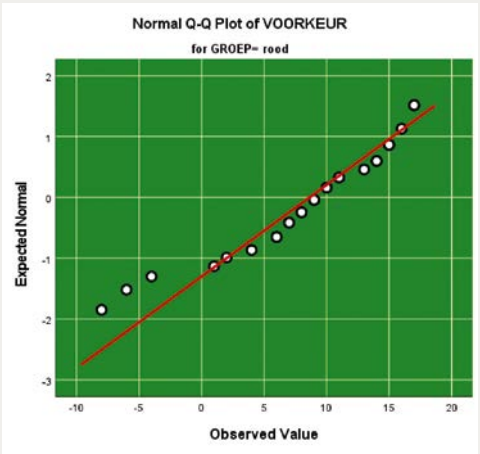












ANOVA

Descriptives

VOORKEUR

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
blauw	30	-5,77	8,873	1,620	-9,08	-2,45	-18	18
neutraal	31	-,87	22,546	4,049	-9,14	7,40	-37	46
rood	30	8,63	6,641	1,212	6,15	11,11	-8	17
Total	91	,65	15,646	1,640	-2,61	3,91	-37	46

ANOVA

VOORKEUR

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3218,930	2	1609,465	7,529	,001
Within Groups	18811,817	88	213,771		
Total	22030,747	90			