Piraten, perziken en p-waarden

UITWERKINGEN VAN DE OPDRACHTEN

VERSIEDATUM: 29-01-2017

HOOFDSTUK 1 GEGEVENS IN KAART

1. A

De verdeling is scheef naar rechts; de hoge score aan de rechterkant (wellicht een uitschieter?) trekt het gemiddelde omhoog, maar niet de mediaan, die iets meer in de buurt van de top zit.

2. A

Aangezien 25% van de dromen 10 minuten of korter duurde, vormt 10 de eerste kwartielgrens: $Q_1 = 10$. De derde kwartielgrens wordt dan ook gevormd door het getal 25: $Q_3 = 25$. Dat betekent dat

$$IQR = Q_3 - Q_1 = 25 - 10 = 15$$
$$1{,}5 * IQR = 1{,}5 * 15 = 22{,}5$$
$$Q_3 + 1{,}5 * IQR = 25 + 22{,}5 = 47{,}5$$

Die 47,5 minuut is de bovengrens voor 'gebruikelijke' droomduur. Aangezien 60 minuten hier nog ruim boven ligt, mogen we deze droom als een uitschieter beschouwen.

3. C

Van verhoudingen (ratio's) zoals bij antwoord A kunnen we alleen spreken bij een ratiovariabele; de getallen op de Likertschaal hebben géén kwantitatieve betekenis! Ook zijn de afstanden tussen opeenvolgende getallen niet zonder meer even groot. Daarmee valt ook B af: we weten niet zeker of het verschil tussen 1 en 2 even groot is als het verschil tussen 2 en 3. Antwoord C kun je wel zeggen bij een ordinale schaal.

4.

a) Zeker, een modus kan altijd. Deze is 2.
b) Een mediaan kan vanaf een ordinale variabele, dus laat maar komen. Tel de balkhoogten: we hebben in totaal 16 scores. De mediaan is dus score nummer $\frac{N+1}{2} = \frac{17}{2} = 8{,}5$ – oftewel het gemiddelde tussen de achtste en de negende score. Weer naar het staafdiagram: de achtste score is een 2 en de negende een 3. Daarmee wordt de mediaan 2,5.
c) Liever niet: de Likertschaal is niet kwantitatief!
d) Noppes. Je zou het gemiddelde nodig hebben en berekeningen moeten doen, hetgeen betekenisloos is bij categorische variabelen.
e) Neuj. De IQR vereist weer een berekening. Was je soms van plan om 3 (deels onlogisch, deels logisch) minus 2 (nogal onlogisch) te doen? En wat had de uitkomst dan betekend? De 50% middelste scores liggen maximaal '1 logica uiteen'? Wat een kolder.

5. D

Met deze opgave wil ik je <u>inhoudelijk</u> laten nadenken over statistische informatie. In mijn persoonlijke ervaringen is deze variabele scheef naar rechts verdeeld: soms ontmoet ik bijna niemand, soms een stuk of vijf individuen en heel af en toe passeren tientallen personen de revue. Mogelijk zijn jouw eigen dromen juist heel drukbezocht, beste lezer, en ben je slechts zelden heel eenzaam (scheef naar links). Of misschien heb je in je dromen doorgaans wel een stuk of drie mensen om je heen, soms minder, soms meer (symmetrisch). Het punt is dat je altijd van tevoren theorieën mag verzinnen, sterker nog: <u>moet</u> verzinnen. Als we de statistiek die we zien niet inhoudelijk kunnen interpreteren, hebben we er niets aan. ☺

6.

a) Eens kijken. De variatie geeft aan hoe personen afwijken van het gemiddelde. We berekenen deze door van alle individuele scores het gemiddelde af te trekken, het verschil te kwadrateren en alle kwadraten vervolgens op te tellen:

$$\sum (X_i - \bar{X})^2$$

Echter, deze nieuwe proefpersoon heeft zijn droom al 1 keer eerder beleefd. Hij of zij scoort daarmee gelijk aan het gemiddelde, en wijkt dus niet af van het gemiddelde. We tellen dus $0^2 = 0$ op bij de huidige variatie.

Bedenk ook eens wat dit inhoudelijk betekent: deze persoon varieert niet van het gemiddelde, en voegt dus ook niets toe aan de mate waarin de deelnemers tezamen variëren.

b) Dit is de gemiddelde variatie per persoon. Aangezien deze ene persoon niet varieert (afwijkt) van $\bar{X}$, zal de gemiddelde afwijking dalen. En inderdaad, om de variantie te krijgen, delen we de variatie nu door een iets groter getal:

$$s_X^2 = \frac{\sum(X_i - \bar{X})^2}{\boldsymbol{N-1}}$$

Niet meer door $16 - 1 = 15$, maar door $17 - 1 = 16$. Daar wordt de variantie iets kleiner van.

c) Dientengevolge wordt ook de standaardafwijking, de wortel uit de variantie, iets kleiner:

$$s_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{\boldsymbol{N-1}}}$$

7. D

Nout en Emma zijn beiden 0,76 standaarddeviaties van het gemiddelde verwijderd; Emma eronder, Nout erboven.

HOOFDSTUK 2 CATEGORISCHE VERBANDEN

1. C

Enkel aan de univariate verdelingen kunnen we niet zien hoe variabelen met elkaar samenhangen, precies vanwege het feit dat de kruistabel dan nog op diverse manieren ingevuld kan worden. Neem bijvoorbeeld deze situatie:

		DROOMTYPE		
		aangenaam of neutraal	nachtmerrie	
TERUGKERENDE DROOM	nee	21 (87,5%)	7 (87,5%)	28 (87,5%)
	ja	3 (12,5%)	1 (12,5%)	4 (12,5%)
		24 (100%)	8 (100%)	32 (100%)

Hier hebben we geen samenhang.
Of deze:

		DROOMTYPE		
		aangenaam of neutraal	nachtmerrie	
TERUGKERENDE DROOM	nee	24 (100%)	4 (50%)	28 (87,5%)
	ja	0 (0%)	4 (50%)	4 (12,5%)
		24 (100%)	8 (100%)	32 (100%)

Hier hebben we een zeer sterke samenhang.

2. B

Antwoord A is geen eerlijke vergelijking, want er waren veel meer aangename of neutrale dromen. In dat geval is het juist merkwaardig dat er even veel terugkerende dromen bij zaten (2) als bij de nachtmerries. Waar het om gaat is of bij beide droomtypen relatief even veel terugkerende dromen zaten. We zijn dus geïnteresseerd in de percentages:

		DROOMTYPE		
		aangenaam of neutraal	nachtmerrie	
TERUGKERENDE DROOM	nee	22 (91,7%)	6 (75%)	28 (87,5%)
	ja	2 (8,3%)	2 (25%)	4 (12,5%)
		24 (100%)	8 (100%)	32 (100%)

Nu zien we dat nachtmerries relatief vaker terugkeerden dan aangename of neutrale dromen. B is juist. C is niet juist: de niet-terugkerende dromen waren inderdaad altijd in de meerderheid, maar bij aangename of neutrale dromen was het percentage 91,7% en bij nachtmerries lag het lager met 75%. Dan is het nog wel degelijk zo dat het percentage terugkerende dromen verandert met het droomtype.

3.

De kolompercentages heb ik al in opgave 2 berekend. De rijpercentages zijn:

		DROOMTYPE		
		aangenaam of neutraal	nachtmerrie	
TERUGKERENDE DROOM	nee	22 (78,6%)	6 (21,4%)	28 (100%)
	ja	2 (50%)	2 (50%)	4 (100%)
		24 (75%)	8 (25%)	32 (100%)

De rijpercentages geven aan: 'Van degenen zonder/met terugkerende droom had zoveel procent bijvoorbeeld een nachtmerrie.'

De kolompercentages geven aan: 'Van degenen met bijvoorbeeld een nachtmerrie had zoveel procent niet/wel een terugkerende droom.'
Op zich zijn beide soorten percentages informatief. Niettemin heb ik een lichte voorkeur voor percentages die het causale verband het best uitdrukken. Volgens mij beïnvloedt het droomtype de kans op een terugkerende droom, en niet andersom. Deze causale richting wordt het best beschreven door de kolompercentages.

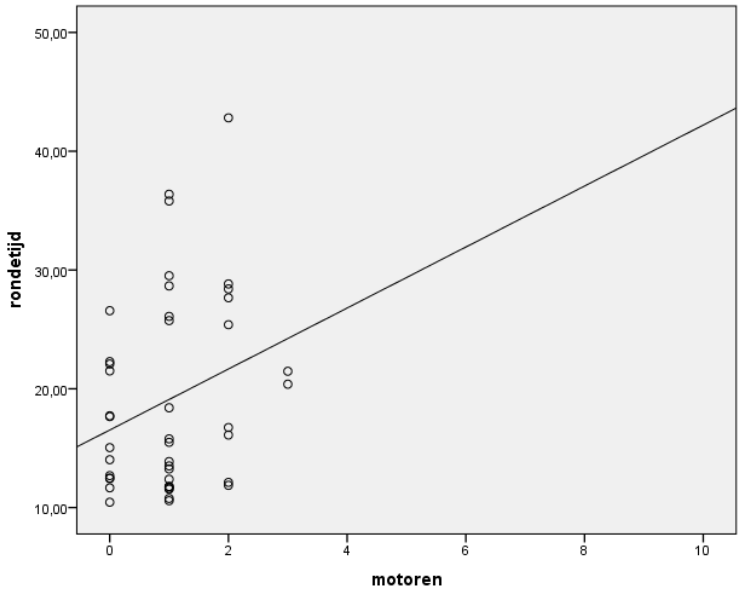
4. C

Dit soort statistiek is misleidend (en dus gevaarlijk): mensen met nachtmerries zijn in groten getale abrupt ontwaakt… maar wie zegt dat dat niet gold voor mensen met aangename of neutrale dromen? Pas als we deze mensen naast degenen met nachtmerries leggen, kunnen we zien of het ene droomtype vaker abrupt ontwaakt dan het andere. Nu we alleen de individuen met nachtmerries bekijken, varieert het droomtype niet; verbanden kunnen echter enkel bestaan tussen variabelen.

HOOFDSTUK 3 KWANTITATIEVE VERBANDEN

1. A

Je kunt nu eenmaal geen halve motor installeren. Het aantal gebruikte motoren is niettemin nog steeds een ratiovariabele en zeker niet ordinaal. Zien we dan vier subgroepen? Welnee: er zou een verborgen variabele moeten zijn die deze subgroepen creëert, maar in feite bestaat elk 'stapeltje' scores gewoon uit coureurs met een bepaald aantal motoren op hun zeepkist (X). Kijk voor de aardigheid eens hieronder: zo zou het spreidingsdiagram eruit hebben gezien als de X-as tot 10 motoren ging.



Ziet dat er niet uit als een heel normaal spreidingsdiagram? ☺

2.

Hoe meer motoren…: GOED. We kunnen zeggen dat er een positieve trend is.
De regressielijn vertoont…: FOUT. De correlatiecoëfficiënt vertelt ons alleen hoe sterk het verband is en dus hoe goed de punten de regressielijn volgen. Over de helling van de lijn zegt hij niets (behalve dat deze positief is).
Motoren hebben een…: FOUT. De voorspelde rondetijd stijgt als de coureur meer motoren gebruikt… Is dat gunstig? Nee! Het duurt langer om de ronde te rijden! De organisator had juist een dalende trend verwacht. Het lijkt erop dat coureurs met motoren sneller de controle kwijtraken over hun voertuig en bijvoorbeeld uit de bocht vliegen.
Behalve het aantal…: GOED. De correlatie is niet 1, dus het aantal motoren is niet het enige wat de rondetijd beïnvloedt. Dat lijkt me logisch. Naast overige eigenschappen van de zeepkist (wegligging, wendbaarheid, remmen) speelt ook de vaardigheid van de coureur een grote rol.

3. B

Zie eventueel de theorie (slotparagraaf). We hebben nu een vrij klein bereik van aantallen motoren onderzocht, en dat drukt de correlatie. Mogelijk zou deze correlatie groter worden als er ook coureurs met nog meer motoren hadden rondgereden. We weten dit overigens niet zeker. Zou de lineaire trend zich wel voortzetten? Het kan ook dat nóg een extra motor op een gegeven moment geen verschil meer maakt voor de rondetijd (en dat de lijn dus afvlakt). Wie de regressielijn doortrekt tot in een ongemeten gebied, is bezig te **extrapoleren**. Extrapolatie is dus altijd een beetje ongewis.

4.

a)

$$b = r_{XY} * \frac{s_Y}{s_X} = 0{,}268 * \frac{8{,}107}{0{,}845} = 2{,}57$$
$$a = \bar{Y} - b\bar{X} = 19{,}10 - 2{,}57 * 1 = 16{,}53$$

Kortom,

$$\hat{Y} = 16{,}53 + 2{,}57X$$

b) De voorspelde rondetijd voor een zeepkist zonder motoren ($X = 0$); deze bedraagt 16,53 minuten.
c) De voorspelde toename van de rondetijd per extra motor; deze bedraagt 2,57 minuten.
d) Geen van beide: om te bepalen hoe sterk de rondetijd afhankelijk is van de hoeveelheid motoren, hebben we de correlatie nodig.

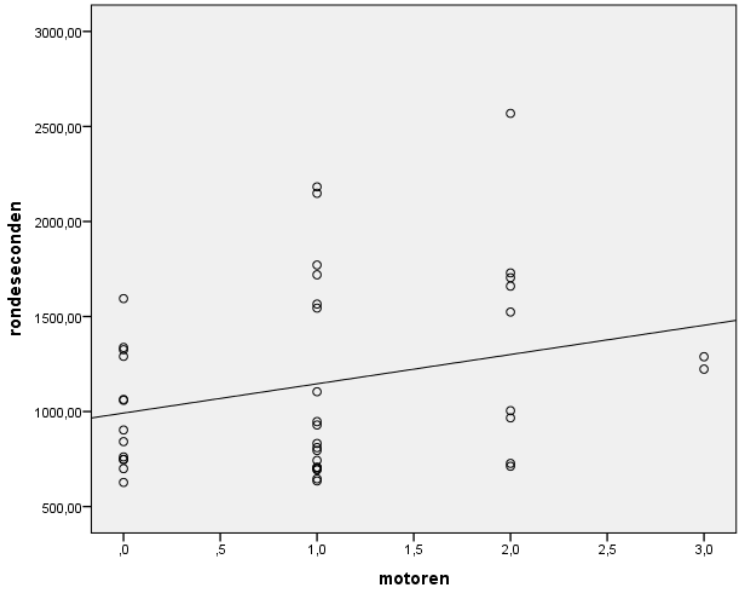
5.
Wat mij betreft wel: ik zie geen andere samenhang meer (zoals een kromlijnige).

6.

$$R^2 = 0{,}268 = 0{,}072$$

Oftewel, de variatie in de rondetijden kan voor 7,2% worden verklaard doordat niet iedere zeepkist even veel motoren had. Dat is niet erg veel: 92,8% van de variatie moet dus worden toegeschreven aan andere factoren, zoals het feit dat de coureurs verschillen in talent. Het is dan ook de vraag of de organisator de verklaring voor de race-uitslag bij motoren moet leggen. Zo niet, dan hoeft het gebruik van motoren volgend jaar ook niet te worden verplicht.

7. B
De organisator heeft de rondetijd getransformeerd: van minuten naar seconden. Het betreft hier een multiplicatie, want alle tijdscores worden nu 60 keer zo groot. Betekent dat dat de scores dichter bij of verder van elkaar komen te liggen? Nee, want ze veranderen allemaal in precies dezelfde mate!



Dus onthoud: als we variabelen lineair transformeren, blijft hun correlatie exact gelijk.

8. D
Geen samenhang betekent een vlakke regressielijn ($b = 0$) en geen correlatie ($r = 0$), dus ook een proportie verklaarde variatie van noppes ($R^2 = 0$). De regressielijn stelt nu gewoon de gemiddelde lijn voor. Zie eventueel het plaatje aan het eind van paragraaf 3.4, onder het kopje *Bijzondere situaties*.

9. C
De twee genoemde variabelen zijn categorisch! Een correlatie mogen we alleen uitrekenen tussen kwantitatieve variabelen.

HOOFDSTUK 4 BASIS KANSREKENING

1.

a)

		OMKOPERIJ		
		nee	ja	
PRIJS	knuffelbeer	12	12	24
	plastic prul	228	18	246
		240	30	270

b) Een kwestie van begrijpend lezen. We zoeken naar

$$P(omkoperij\ en\ knuffelbeer) = \frac{12}{270} = 0{,}044$$

c) Dit is niet hetzelfde als bij vraag b! Het gaat nu om een voorwaardelijke kans: de kans op omkoperij, als het kind een knuffelbeer gewonnen heeft. Schrijf de gevraagde kans correct op en het komt goed:

$$P(omkoperij|knuffelbeer) = \frac{12}{24} = 0{,}5$$

d) Nee, zeker niet: er zijn immers 12 kinderen met zowel een knuffelbeer als omkopende ouders.

e) Eens kijken. We hebben statistische onafhankelijkheid indien

$$P(A) = P(A|B)$$

Voldoen de data hieraan?

$$P(omkoperij) = \frac{30}{270} = 0{,}111$$

Echter,

$$P(omkoperij|knuffelbeer) = \frac{12}{24} = 0{,}5$$

… zoals we in opgave b al berekenden. De kans op omkopende ouders stijgt dus flink indien we te maken hebben met een kind dat een knuffelbeer gewonnen heeft. De gebeurtenis 'omkoperij' is sterk afhankelijk van de gebeurtenis 'knuffelbeer'.

Ook een goede oplossing:

$$P(knuffelbeer) = \frac{24}{270} = 0{,}089$$

Echter,

$$P(knuffelbeer|omkoperij) = \frac{12}{30} = 0{,}4$$

Ja, deze laatste kans stond al in de introductie… met andere woorden, de introductie vertelde ons al dat de kans op een knuffelbeer deels afhankelijk is van de vraag of de ouders de eigenaar omkopen.

2.

Versie 2015 (handboek)

a) Dit is een gebeurtenis: we hadden ook twee andere kinderen, een andere EU, kunnen trekken die beiden een plastic prul hebben gewonnen! De elementaire uitkomst is de getrokken steekproef, bijvoorbeeld 'Marieke, Pip'.

b) Simpel.

$$P(plastic\ prul\ en\ plastic\ prul) = \frac{246}{270} * \frac{246}{270} = 0{,}830$$

c) Iets minder simpel: let erop dat er twee volgorden mogelijk zijn (zie de laatste theorieparagraaf)! Of het eerste kind wint een knuffelbeer en het andere een plastic prul, of andersom.

$$P(knuffelbeer\ en\ plastic\ prul) = \frac{24}{270} * \frac{246}{270} * 2 = 0{,}162$$

d) Dat wordt lastig… want dan zijn de twee gebeurtenissen niet meer onafhankelijk van elkaar! Stel dat het jongetje een knuffelbeer wint? Dan zullen zijn ouders misschien meer geneigd zijn de eigenaar om te kopen, opdat zijn zusje meer kans zal maken om ook een knuffelbeer te winnen. We weten echter niet zeker wat de ouders gaan doen. Daarom kunnen we niet de echte kans bepalen dat het zusje alsnog een plastic prul zal winnen.

Versie 2016

a) Simpel.

$$P(plastic\ prul\ en\ plastic\ prul) = \frac{246}{270} * \frac{246}{270} = 0{,}830$$

b) Iets minder simpel: let erop dat er twee volgorden mogelijk zijn (zie de laatste theorieparagraaf)! Of het eerste kind wint een knuffelbeer en het andere een plastic prul, of andersom.

$$P(knuffelbeer\ en\ plastic\ prul) = \frac{24}{270} * \frac{246}{270} * 2 = 0{,}162$$

c) Wederom simpel.

$$P(knuffelbeer\ en\ knuffelbeer) = \frac{24}{270} * \frac{24}{270} = 0{,}008$$

d) Dit is een gebeurtenis: een verzameling van meerdere elementaire uitkomsten. Je kunt allerlei steekproeven trekken van twee kinderen (één steekproef vormt één EU) die beiden een plastic prul hebben gewonnen! Zojuist hebben we uitgerekend dat dit gebeurt bij 83,0% van alle EU's.
Overigens klopt de kansverdeling, want opgeteld zijn de drie kansen gelijk aan 1. ☺

e) Dat wordt lastig… want dan zijn de twee gebeurtenissen niet meer onafhankelijk van elkaar! Stel dat het jongetje een knuffelbeer wint? Dan zullen zijn ouders misschien meer geneigd zijn de eigenaar om te kopen, opdat zijn zusje meer kans zal maken om ook een knuffelbeer te winnen. We weten echter niet zeker wat de ouders gaan doen. Daarom kunnen we niet de echte kans bepalen dat het zusje bijvoorbeeld alsnog een plastic prul zal winnen.
Om deze reden gaan bijna alle statistische technieken in *Piraten, perziken en p-waarden* ervan uit dat alle proefpersonen onafhankelijk zijn. Let hierop als je zelf een keer proefpersonen gaat testen!

3. C

A is de populatie; B is één elementaire uitkomst. C bevat alle mogelijke elementaire uitkomsten; een EU is bij steekproeftrekking immers altijd de complete steekproef.

4. B

Herhaling van de definitie: een toevalsvariabele is een grootheid die door toeval verschillende waarden zal aannemen bij herhaling van het kansexperiment.
De omvang van de steekproef ligt vast – we bepalen die zelf – en is dan ook niet toevallig. Als we de omvang veranderen, veranderen we het kansexperiment.
Het aantal ouders dat op deze dag de eigenaar omkocht (namelijk 30) is een **populatie**waarde (een voorproefje op hoofdstuk 5): ook dit aantal ligt vast en zal niet veranderen, hoe vaak we ook steekproeven van 10 kinderen blijven trekken uit de kinderen die vandaag touwtje hebben getrokken.
Echter, het aantal kinderen in de steekproef dat een knuffelbeer wint zal anders zijn bij elke herhaling van het kansexperiment. De ene keer trekken we 0 van zulke geluksvogels, de volgende keer 1, soms zelfs 2 enzovoort.

5.

a) De verwachtingswaarde(!) is

$$E(X) = N * p = 20 * 0{,}17 = 3{,}4$$

Gemiddeld zullen we dus 3,4 brakende passagiers trekken in een steekproef van 20 personen.

b) De variantie kunnen we berekenen met de formule

$$s_X^2 = N * p * (1 - p) = 20 * 0{,}17 * 0{,}83 = 2{,}82$$

De standaardafwijking is dus

$$s_X = \sqrt{2{,}82} = 1{,}68$$

Oftewel, het aantal brakende passagiers zal in de steekproef gemiddeld met 1,68 afwijken van de verwachtingswaarde.

6. C

Verzamel eerst gegevens, is mijn advies. Er staan er genoeg in de opgave, maar we moeten ze wat overzichtelijker zien op te schrijven. Sowieso is de kans op een brakend persoon

$$P(overgeven) = 0{,}17$$

En die andere percentages, wat voor kansen zijn dat? Let erop dat je ze goed opschrijft!

$$P(volwassene|overgeven) = 0{,}70$$
$$P(kind\ of\ puber|overgeven) = 0{,}30$$

Immers, indien de persoon tot de brakende passagiers behoort (voorwaarde), is er een zekere kans dat hij of zij volwassen is of niet.
Nu, wat zeggen deze gegevens? Is de kans op een kind gelijk aan 0,3? Nee: dat is alleen zo indien de persoon in kwestie moest braken. We weten niet of de algemene kans op een kind hetzelfde is.
Is $P(overgeven|volwassene)$ dan gelijk aan 0,7? Ook niet: wie voor antwoord B kiest, heeft die 70% in de introductie 'verkeerdom' geïnterpreteerd.
Kunnen we misschien komen aan de kans op een volwassen passagier die ook moet braken? Met de productregel soms?

$$P(volwassene\ en\ overgeven) = P(volwassene) * P(overgeven|volwassene)$$

Nee, zo niet…

$$\boldsymbol{P(\ overgeven\ en\ volwassene)} = P(overgeven) * P(volwassene|overgeven) = 0{,}17 * 0{,}70 = \mathbf{0,119}$$

… maar zo wel! ☺

HOOFDSTUK 5 KANSVERDELINGEN

1.

a) De 68-95-99,7-vuistregel vertelt ons dat 95% van de worstplakjes maximaal 2 standaarddeviaties van het gemiddelde afwijkt.

$$9 - 2 * 3{,}5 = 2$$
$$9 + 2 * 3{,}5 = 16$$

Antwoord: 95% van de plakjes is tussen de 2 en 16 millimeter dik.

b) Naast de plakjes worst in opgave a is er nog 5% over. Een normale verdeling is symmetrisch, dus die 5% is gelijkmatig verdeeld over het linker- en rechteruiteinde van de verdeling. (Schets dit!) 2,5% van de plakjes is dunner dan 2 millimeter en 2,5% is dikker dan 16 millimeter. Kortom, de 2,5% dikste plakken zijn 16 millimeter of dikker.

c) Voor deze vraag is de vuistregel niet geschikt meer. We zullen een z-score moeten berekenen. Dit is gewoon een z-score betreffende de populatieverdeling:

$$Z = \frac{X - \mu}{\sigma} = \frac{12 - 9}{3{,}5} = 0{,}86$$

Uit de z-tabel blijkt dat

$$P(Z > 0{,}86) = 1 - 0{,}8051 = 0{,}1949$$

Dus Luca loopt bij 19,49% van zijn plakjes het risico om op zijn kop te krijgen. Poeh, ik zou er maar zenuwachtig van worden…

2. A

Dit percentage is een schatter: een steekproefwaarde, verkregen uit een aselecte steekproef. Het is een schatting van de populatieparameter, het percentage van álle plakjes die te dik gesneden zijn. De steekproef is uiteraard de 10 getrokken plakjes zelf, niet hun dikte.

Populatie: ALLE PLAKJES CHORIZO	**Steekproef:** DE 10 GEMETEN PLAKJES CHORIZO
Parameter: PERCENTAGE TE DIKKE PLAKJES IN DE POPULATIE (40%?)	**Schatter:** PERCENTAGE TE DIKKE PLAKJES IN DE STEEKPROEF (40%)

3. C

Zoek indien nodig de definities van zuiverheid en efficiëntie nog eens op. De steekproefgemiddelden zijn zuiver, want ze komen uit een aselecte steekproef; als we oneindig vaak een nieuw gemiddelde bepaalden, zouden we gemiddeld op 9 millimeter uitkomen (μ). Wel zijn de twee steekproefresultaten van 7,5 en 10,6 millimeter nog geen schoten in de roos: ze ontberen nauwkeurigheid. We mogen ze dan ook inefficiënt noemen.

4.

a) Deze steekproef is te klein ($N = 7$); de verdelingsvorm en de schatters zijn niet erg betrouwbaar (efficiënt). We kunnen beter geen uitspraken doen over de populatieverdeling en de steekproevenverdeling van het gemiddelde.

b) Deze steekproef is erg groot ($N = 70$); de vorm van de verdeling van steekproefscores benadert waarschijnlijk de populatieverdeling, die dus vast eveneens symmetrisch en eentoppig is[1]. Het is aannemelijk dat het populatiegemiddelde μ_X ongeveer gelijk is aan 6 centimeter, en de standaarddeviatie σ_X gelijk aan 1,5 centimeter. De steekproevenverdeling van het gemiddelde is bij benadering normaal, net zoals de populatieverdeling. De verwachtingswaarde is gelijk aan het populatiegemiddelde, namelijk ongeveer 6, en de standaardfout is $\sigma_{\bar{X}} = {}^{\sigma_X}/_{\sqrt{N}} \approx {}^{1{,}5}/_{\sqrt{70}} = 0{,}18$. We zien aan deze standaardfout hoe efficiënt het steekproefgemiddelde is: gemiddeld zal het met slechts 0,18 centimeter afwijken van het populatiegemiddelde.

c) Ook deze steekproef is groot ($N = 70$); groot genoeg om uitspraken te doen over de populatieverdeling en de steekproevenverdeling van het gemiddelde. Opnieuw benadert de verdeling van steekproefscores waarschijnlijk de populatieverdeling, die dus vast eentoppig is en scheef naar links. Het populatiegemiddelde μ_X zal ongeveer gelijk zijn aan 2 en de standaarddeviatie σ_X ongeveer gelijk aan 0,5. De steekproevenverdeling is echter niet scheef zoals de populatieverdeling, maar vanwege de grote N ongeveer normaal (centrale limietstelling)! Verwachtingswaarde en standaardfout: $\mu_{\bar{X}} \approx 2$, $\sigma_{\bar{X}} \approx {}^{0{,}5}/_{\sqrt{70}} = 0{,}060$. Zeer kleine standaardfout! ☺

[1] Dat is niet zonder meer hetzelfde als 'normaal', maar voor nu vind ik het niet erg als je een symmetrische verdeling altijd als normaal beschouwt. In hoofdstuk 10 gaan we voor het eerst wat extra nuances maken. ☺

d) Enkel de grootte van de steekproef doet ertoe. De steekproevenverdeling van $\bar{X}$ beschrijft welke $\bar{X}$'en Carlo hoe vaak zou krijgen, als hij oneindig veel steekproeven trok. Deze steekproevenverdeling is dus een theoretische gedachte: deze ene steekproef vormt een van de resultaten die Carlo had kunnen verkrijgen. Al die mogelijke resultaten zijn vanwege de grote N ongeveer normaal verdeeld.

e) Daar lijkt het op dit moment wel op: het gemiddelde waterpeil is naar schatting gezakt van 6 naar 2 centimeter. Zeker weten kunnen we het echter niet, aangezien het steekproefschattingen blijven. Kunnen we dan nooit uitsluiten dat deze daling puur toevallig is? Nou… vanaf hoofdstuk 6 gaan we een poging doen!

5.

a) 1,6 gram: het gaat om de standaardafwijking van de verdeling van steekproefscores. De standaarddeviatie van een grote steekproef zal waarschijnlijk sterk lijken op de populatiestandaardafwijking.

b) Het gaat om de kans op een bepaald steekproefgemiddelde. Dit is dus een kans onder de steekproevenverdeling. Schets deze!
De z-score is $Z = \frac{32-33}{1{,}6/\sqrt{30}} = -3{,}42$
Uit de z-tabel volgt: $P(Z < -3{,}42) = 0{,}0003$
Kortom, de kans dat we het gemiddelde gewicht van de populatie stroopwafels ook maar 1 grammetje te laag inschatten, is bijzonder klein! Een steekproef van 30 wafels is zeer betrouwbaar.

c) Dat klopt, maar we berekenen de kans op een steekproefgemiddelde van 32 of lager onder de steekproevenverdeling – en die is vanwege de grote N (30) wél ongeveer normaal, ondanks de scheve populatieverdeling.

6. B

Schets de z-verdeling: volgens de opgave mogen we ervan uitgaan dat deze symmetrisch is en het gemiddelde is in elk geval 0. De z-score van Luca's wafel ligt dan natuurlijk links van het gemiddelde, oftewel lager. Deze wafel is lichter dan gemiddeld.[2]
Meer dan 50% van de scores ligt rechts van Luca's z-score, dus meer dan de helft van de stroopwafels (15) is zwaarder dan die van Luca. Mochten we willen opzoeken hoeveel wafels er zwaarder waren, dan kan dit in de z-tabel (mits we de verdeling als normaal beschouwen): $P(Z < -0{,}15) = 0{,}4404$. Luca's wafel behoort dus grofweg tot de lichtste 44%, maar zeker niet tot de lichtste 15%! Jammer, nu heeft hij geen excuus om er nog een te pakken…

[2] Geldt ook bij een scheve verdeling.

HOOFDSTUK 6 HYPOTHESETOETSING

1.

a) $Z = \frac{6{,}4-5}{3/\sqrt{22}} = 2{,}19$

b) $p = P(Z > 2{,}19) = 1 - P(Z < 1{,}58) = 1 - 0{,}9857 = 0{,}0143$
Significant (bij $\alpha = 0{,}05$)!

c) $p = P(Z < 2{,}19) = 0{,}9857$
Totaal niet significant…

d) $p = 2 * P(Z > 2{,}19) = 2 * 0{,}0143 = 0{,}0286$
Ondanks de verdubbelde p-waarde nog steeds significant.

2. A

Het steekproefgemiddelde vormt het centrum van het betrouwbaarheidsinterval; dit steekproefgemiddelde zal bij elke nieuwe steekproef veranderen, dus het centrum van het betrouwbaarheidsinterval verandert mee. De foutmarge daarentegen bestaat uit de kritieke z-waarde en de standaardfout; deze zijn beide constant. De foutmarge zal dan ook niet veranderen.

3.

a) Hij zou de betrouwbaarheid kunnen verlagen; dan daalt de kritieke z-score en daarmee vermindert de breedte. Echter, in dat geval weet hij minder zeker dat het populatiegemiddelde zich in zijn interval bevindt. Niet zo'n goed idee dus. Beter idee: de steekproef vergroten! Dan daalt immers de standaardfout en ook daarmee vermindert de breedte.

b) Hij zou de kritieke z-waarde moeten ophogen. Daarmee stijgt de betrouwbaarheid. De consequentie is wel dat het interval breder wordt en de producent met wat minder nauwkeurigheid kan zeggen wat het populatiegemiddelde zal zijn.

4.

a) $H_0: \mu = 32$
$H_A: \mu > 32$

b) Deze vraag is een stukje herhaling van hoofdstuk 5. De hypothesetoets kan alleen uitgevoerd worden als de populatieverdeling van de stroopwafelgewichten <u>normaal</u> is. De steekproevenverdeling moet namelijk normaal zijn, omdat de kansen die we in de z-tabel opzoeken op een normale verdeling zijn gebaseerd. De centrale limietstelling geldt niet, omdat de steekproef erg klein is. De steekproevenverdeling is daarom alleen nog normaal indien ook de populatieverdeling dit is. (Eigenlijk gaat die vlieger niet op: kijk maar terug naar opgave 5 van hoofdstuk 5.)

c) Tip: schets de steekproevenverdeling!
Cohens d is naar schatting $\hat{d} = \frac{32{,}8-32}{1{,}6} = 0{,}5$. Een medium effect dus, volgens de richtlijn.
De z-score is $Z = \frac{32{,}8-32}{1{,}6/\sqrt{10}} = 1{,}58$
Uit de z-tabel volgt: $p = P(Z > 1{,}58) = 1 - P(Z < 1{,}58) = 1 - 0{,}9429 = 0{,}0571$
Het toetsresultaat is dus nét niet significant. Dit komt waarschijnlijk doordat de steekproef te klein is: de toets heeft dan weinig *power*. (Verderop moet je zelf de *power* berekenen.) Een iets grotere steekproef – met een kleinere standaardfout – zou waarschijnlijk wel een significant resultaat hebben opgeleverd.

d) De formule voor het betrouwbaarheidsinterval is $\bar{X} \pm Z^* \sigma/\sqrt{N}$. Invullen geeft:

$$32{,}8 - 1{,}645 * 1{,}6/\sqrt{10} = 32{,}8 - 0{,}83$$
$$32{,}8 + 1{,}645 * 1{,}6/\sqrt{10} = 32{,}8 + 0{,}83$$
$$[31{,}97; 33{,}63]$$

Oftewel, het populatiegemiddelde ligt waarschijnlijk (met 90% zekerheid) tussen de gewichten 31,97 en 33,63 gram.

5. C

Zeer belangrijk: de p-waarde vertelt ons niet de kans dat de nulhypothese waar is! Dat zou weliswaar gemakkelijker en intuïtiever zijn geweest, maar zoiets is niet mogelijk: de nulhypothese van dit onderzoek is óf waar, óf niet waar – daar is geen kans mee gemoeid. De p-waarde gaat ervan uit dat de nulhypothese waar is, en vertelt ons wat <u>in dat geval</u> de (voorwaardelijke) kans is op het steekproefgemiddelde dat we hebben gevonden. Waarna wij zeggen: 'Oké, dit steekproefgemiddelde zou niet zo opvallend zijn als de nulhypothese juist is' (de kans is groter dan – standaard – 5%), <u>of</u> 'Ha, maar dan zou mijn steekproefgemiddelde wel héél bijzonder zijn als de nulhypothese juist is!' (de kans is kleiner dan of gelijk aan 5%). In het laatste geval verwerpen we de nulhypothese.

6.

a) Nee: alleen de kans op een Type II-fout zal dalen met een grotere steekproef. De kans op een Type I-fout is altijd gelijk aan het significantieniveau α. De fabrikant zou dan ook dit significantieniveau moeten verkleinen als hij bang voor Type I-fouten is. Het nadeel: dit kost *power*! Verklein in de tekening van paragraaf 6.5 maar eens het significantieoppervlak. De kritieke grens schuift op naar rechts en het groene oppervlak onder de werkelijke steekproevenverdeling (de *power*) wordt kleiner. Tot zover de wiskunde. Is dit ook conceptueel logisch? Ja: bij een kleinere α maak je het jezelf <u>moeilijker</u> de nulhypothese te verwerpen, en dus verklein je de kans dat het je terecht lukt.

b) Niet 5%: deze kans is gelijk aan 0. De onderzoeker verwerpt de nulhypothese niet, dus hij kan deze niet onterecht verwerpen. Strikvraag! MWOEAHAHA!

c) Gebruik ook mijn uitwerking achter in de theorie (in het handboek aangeduid als het hulpmiddel van de geheimzinnige Snorro).
Teken de steekproevenverdeling volgens de nulhypothese en de werkelijke steekproevenverdeling naast elkaar; laat je hierbij inspireren door de afbeelding in paragraaf 6.5.
Het kritieke steekproefgemiddelde is het steekproefgemiddelde waarbij de nulhypothese $\mu = 32$ zou worden verworpen. Dit gemiddelde heeft een p-waarde van 0,05.
Uit de z- of de t-tabel blijkt de kritieke z-score ongeveer gelijk aan 1,645 (zoek de z-score met een p-waarde van 0,05). $Z^* = \frac{\bar{X}^* - \mu}{\sigma/\sqrt{N}}$ geeft dat $1{,}645 = \frac{\bar{X}^* - 32}{1{,}6/\sqrt{10}}$. Als we deze vergelijking oplossen, krijgen we $\bar{X}^* = 1{,}645 * 1{,}6/\sqrt{10} + 32 = 32{,}83$.

d) De nulhypothese wordt dus verworpen als $\bar{X}$ groter is dan 32,83; de nulhypothese wordt <u>niet</u> verworpen als $\bar{X}$ <u>kleiner</u> is dan 32,83. Wat is nu de <u>werkelijke</u> kans dat $\bar{X}$ kleiner is dan 32,83? Dit is het <u>rode</u> oppervlak onder de werkelijke steekproevenverdeling. We kunnen dit op de ouderwetse manier berekenen:
De z-score is $Z = \frac{32{,}83 - 33}{1{,}6/\sqrt{10}} = -0{,}34$
Uit de z-tabel blijkt dat $\boldsymbol{P(Type\ II\text{-}fout)} = P(Z < -0{,}34) = \mathbf{0{,}3669}$.
Dit is een aanzienlijke kans!

e) De *power* is gelijk aan het groene oppervlak onder de werkelijke steekproevenverdeling: $1 - P(Type\ II\text{-}fout) = 1 - 0{,}3669 = \mathbf{0{,}6331}$. Dit is niet zo'n grote *power*. Een grotere steekproef kan dit probleem verhelpen.

HOOFDSTUK 7-9 T-TOETSEN

1.

a) Vergis je niet: dit is een onderzoek naar 1 populatie. Van de Nova Fyra was de populatieverdeling al bekend, dus het zou nauwkeurigheid (en dus *power*) kosten om ook uit deze oude modellen een steekproef te trekken. Research and Development trok slechts één steekproef van 30 Photons; geschikt was daarom een *one sample t-test*.

b) $H_0: \mu = 1000$
$H_A: \mu > 1000$

c) Cohens d was naar schatting $\hat{d} = \frac{1041-1000}{38} = 1{,}08$. Een groot effect dus!
De t-score was $T = \frac{1041-1000}{38/\sqrt{30}} = 5{,}909$.
Teken eventueel de t-verdeling, met het gemiddelde (0) en de gevonden t-score.
Het aantal vrijheidsgraden bedroeg $30 - 1 = 29$.
Uit de t-tabel blijkt dat $p = P(T > 5{,}909) \ll 0{,}0005$.
Dit was knettersignificant: de nulhypothese kon keihard verworpen worden. Goed nieuws voor de Sterrenvloot: de Photon was sneller dan de Nova Fyra!

d) De formule voor het betrouwbaarheidsinterval is $\bar{X} \pm T^* {}^{S}/_{\sqrt{N}}$. Invullen geeft (afgerond op gehele kilometers per uur):

$$1041 - 2{,}756 * {}^{38}/_{\sqrt{30}} = 1041 - 19$$
$$1041 + 2{,}756 * {}^{38}/_{\sqrt{30}} = 1041 + 19$$
$$[1022; 1060]$$

Oftewel, het populatiegemiddelde lag zeer waarschijnlijk (met 99% zekerheid) tussen de snelheden 1022 en 1060 kilometer per uur.

2. A

Dit was een *paired samples t-test*, dus de gemeten groepen moesten afhankelijk van elkaar zijn. Bij A was er sprake van herhaalde metingen; bij B werden er verschillende gevechtsschepen met elkaar vergeleken. In theorie had het onderzoeksteam individuele Photons en Z-Wings mogelijk gematcht, maar ik zou niet echt weten waarop. Zolang we niet meer informatie hebben, ligt het voor de hand dat de Photons en de Z-Wings onafhankelijke groepen vormden. Voor het onderzoek bij B was daarom een *independent samples t-test* nodig.

3. B

De *paired samples t-test* komt rekenkundig overeen met een *one sample t-test*, uitgevoerd over de verschilscores tussen twee metingen.

4.

a) Jazeker: de waarnemer vergeleek twee onafhankelijke steekproeven.

b) Variabele kwantitatief: voldaan.
Onafhankelijke groepen: voldaan.
Normaliteit: wat mij betreft zien de histogrammen (de verdelingen van steekproefscores) er prima uit, beste lezer. Mocht jij toch je twijfels hebben over deze assumptie, dan waren de steekproeven groot genoeg om voor lichte schending te corrigeren.
Gelijke varianties: hieraan was waarschijnlijk voldaan, want de vuistregel geeft geen reden tot paniek $\left(\frac{40{,}637}{35{,}760} = 1{,}14 < 2\right)$ en *Levene's Test* is geenszins significant ($p = 0{,}715$).
Conclusie: geen schendingen die problematisch zouden zijn.

c) We mogen uitgaan van gelijke varianties (zie vraag b), dus we hebben $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ nodig.

$$s_p = \sqrt{\frac{21 * 35{,}760^2 + 18 * 40{,}637^2}{21 + 18}} = 38{,}089$$
$$s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 38{,}089\sqrt{\frac{1}{22} + \frac{1}{19}} = \mathbf{11{,}929}$$

Dit getal klopt met de *Standard Error of the Difference* in de rij *Equal variances assumed*.

d) Het juiste aantal vrijheidsgraden is $\boldsymbol{df} = n_1 + n_2 - 2 = 22 + 19 - 2 = \mathbf{39}$. Ook dat klopt met de uitvoer in de rij *Equal variances assumed*.

5. B

Hij had namelijk zijn toevlucht kunnen nemen tot… het betrouwbaarheidsinterval! De waarde 0 ligt niet in dit interval (overigens noch bij *Equal variances assumed* noch bij *Equal variances not assumed*, hoewel de eerste keuze eigenlijk de juiste is). 0 is dus geen aannemelijk verschil tussen de twee populatiegemiddelden. Het significantieniveau bij een 95%-betrouwbaarheidsinterval is gelijk aan $100\% - 95\% = 5\%$. De nulhypothese $H_0: \mu_{training} = \mu_{controle}$, ook wel $H_0: \mu_{training} - \mu_{controle} = 0$, moet bij dit significantieniveau dan ook verworpen worden.

6. C

De nulhypothese kunnen we verwerpen: de training had aantoonbaar een effect, dus B valt af. De vraag is nu, werkte het effect ook in de gewenste richting? Kijken we nog eens naar de steekproefscores *(Group Statistics)*, dan zien we dat de gemiddelde verbetering bij de getrainde piloten juist stukken lager lag: ze gingen gemiddeld flink achteruit! De piloten op de wachtlijst scoorden veel hoger (namelijk ongeveer 0; ze veranderden gemiddeld nauwelijks, wat logisch is, aangezien er niet met hen werd geëxperimenteerd). Dit verschil tussen de trainings- en de controlegroep bleek significant. De conclusie van de waarnemer was dus antwoord C… 😐

HOOFDSTUK 10 EENWEG-ANOVA

Opdrachten voor perziken

1.

a) $H_0: \mu_{1\ (Fungus\ Fantasticus)} = \mu_{2\ (Rainbow\ Spore)} = \mu_{3\ (Psycho\ Classic)}$
$H_A: niet\ alle\ \mu's\ zijn\ gelijk$

b) Eerst moeten we de gemiddelden van alle groepen uitrekenen, en het grote gemiddelde erbij:

$$\bar{Y}_1 = \frac{2+7+6+1}{4} = 4$$
$$\bar{Y}_2 = \frac{10+6+10+10}{4} = 9$$
$$\bar{Y}_3 = \frac{2+2+9+7}{4} = 5$$
$$\bar{Y} = \frac{2+7+6+1+10+8+8+10+4+6+5+5}{12} = 6$$

Dan kunnen we nu beginnen met de kwadratensommen. Zie je niet wat er in deze berekeningen gebeurt? Bekijk de theorie nog eens!

$$\boldsymbol{SS(Total)} = \sum_{ij} \left(\boldsymbol{Y_{ij} - \bar{Y}}\right)^2 =$$
$$(2-6)^2 + (7-6)^2 + (6-6)^2 + (1-6)^2 + (10-6)^2 + (6-6)^2 +$$
$$(10-6)^2 + (10-6)^2 + (2-6)^2 + (2-6)^2 + (9-6)^2 + (7-6)^2 =$$
$$16 + 1 + 0 + 25 + 16 + 0 + 16 + 16 + 16 + 16 + 9 + 1 =$$
$$\mathbf{132}$$

$$\boldsymbol{SS(Between)} = \sum_{ij} (\boldsymbol{\bar{Y}_i - \bar{Y}})^2 =$$
$$(4-6)^2 + (4-6)^2 + (4-6)^2 + (4-6)^2 +$$
$$+(9-6)^2 + (9-6)^2 + (9-6)^2 + (9-6)^2 +$$
$$(5-6)^2 + (5-6)^2 + (5-6)^2 + (5-6)^2$$

Let erop dat we <u>voor ieder individu</u> het verschil tellen tussen zijn of haar groepsgemiddelde en het algemene gemiddelde! Maar ja, aangezien dat verschil voor iedereen die in dezelfde groep zit gelijk is, kunnen we het ook gewoon per groep berekenen en vermenigvuldigen met het aantal personen. (Merk op dat dan de j, voor de proefpersoon, onder het somteken verdwijnt.)

$$\boldsymbol{SS(Between)} = \sum_{i} \boldsymbol{n_i} * (\boldsymbol{\bar{Y}_i - \bar{Y}})^2 =$$
$$4 * (4-6)^2 + 4 * (9-6)^2 + 4 * (5-6)^2 =$$
$$4 * 4 + 4 * 9 + 4 * 1 =$$
$$\mathbf{56}$$

$$\boldsymbol{SS(Within)} = \sum_{ij} \left(\boldsymbol{Y_{ij} - \bar{Y}_i}\right)^2 =$$
$$(2-4)^2 + (7-4)^2 + (6-4)^2 + (1-4)^2 + (10-9)^2 + (6-9)^2 +$$
$$(10-9)^2 + (10-9)^2 + (2-5)^2 + (2-5)^2 + (9-5)^2 + (7-5)^2 =$$
$$4 + 9 + 4 + 9 + 1 + 9 + 1 + 1 + 9 + 9 + 16 + 4 =$$
$$\mathbf{76}$$

Ziezo, het ergste ligt achter ons. Nu de ANOVA-tabel. Zie de theorie voor de rekenregels. De p-waarde zoeken we op in de appendix.

	Sum of Squares	df	Mean Square	F	p
Between Groups	56	2	28	3,32	⟨ 0,05 ; 0,10 ⟩
Within Groups	76	9	8,44		
Total	132	11			

c) Conclusie: niet significant. We kunnen niet aantonen dat enigerlei paddenstoel zorgt voor gemiddeld prettigere hallucinaties!

2. A

Minder proefpersonen betekent minder *power* (zie hoofdstuk 6) en dat is een aannemelijke verklaring voor de niet-significante ANOVA in opgave 1. Waren er assumpties geschonden? Dat lijkt me onwaarschijnlijk. Assumpties gaan

altijd over de populatie, dus met de SPSS-uitvoer van de grote steekproef kunnen we controleren of er voor beide ANOVA's aan voldaan is. De histogrammen suggereren vrij symmetrisch verdeelde populaties (ook de Rainbow Spore-groep; maak de meest rechtse balk 1 à 2 personen korter en het is al bijna symmetrisch!). Mocht er niet voldaan zijn aan gelijke varianties, dan had dat voor beide ANOVA's niets uitgemaakt, want alle groepen zijn steeds even groot. Ten slotte is de suggestie van een Bonferroni-correctie flauwekul: die passen we alleen toe op de paarsgewijze vergelijkingen, niet op de ANOVA zelf. Zie de theorie als je niet meer weet waar de Bonferroni-correctie voor bedoeld is.

3. B

Als Bonferroni niet naar ons toe komt, moeten wij maar naar Bonferroni gaan – oftewel, zelf corrigeren. Dat kan prima met de hand! We kunnen het significantieniveau α van elke toets door 3 delen (er zijn immers drie toetsen) of alle p-waarden verdrievoudigen. In dit geval verandert dit niets aan het lijstje met significante resultaten: Rainbow Spore verschilt significant van zowel Fungus Fantasticus als Psycho Classic en dat was het.
Doet Rainbow Spore het nu ook beter dan de andere twee paddenstoelen, of juist slechter? Om deze reden hintte ik naar de *Descriptives*-tabel. Daar zien we dat de testers van Rainbow Spore het hoogste gemiddelde gaven aan hun tripervaring. In de paarsgewijze vergelijkingen is vervolgens gebleken dat dit gemiddelde ook significant hoger ligt dan de andere twee.

4. B

De ANOVA is bepaald niet significant ($p = 0{,}214$), dus we kunnen de nulhypothese niet verwerpen; we krijgen niet aangetoond dat het type paddenstoel (de behandeling) een effect heeft op de kwaliteit van de trip. In dat geval verschilt die kwaliteit tussen individuen enkel door restfactoren.

5. D

Zie de theorie voor details. Ook *MS(Between)* kan hier goed een zuivere schatter zijn van de errorvariantie, aangezien het er niet op lijkt dat de nulhypothese onwaar is. De groepsgemiddelden verschillen in dat geval enkel door erroreffecten.

6. A

De nulhypothese kan worden verworpen als *MS(Between)* opvallend veel groter is dan *MS(Within)*. Immers, *MS(Between)* is een zuivere schatter van de errorvariantie in de populatie, plus de variantie ten gevolge van het groepseffect; terwijl *MS(Within)* een zuivere schatter is van enkel de errorvariantie. Als *MS(Between)* opvallend veel groter is, suggereert dat dat zulke variantie ten gevolge van het groepseffect echt bestaat. De vraag is: is het huidige resultaat opvallend genoeg? Dat blijkt niet zo te zijn, want de ANOVA is niet significant. Een *MS(Between)* die iets meer dan anderhalf keer zo groot is als *MS(Within)* (dit is de F-ratio: 1,594), blijkt nog best vaak voor te komen als de nulhypothese waar is: 21,4% van de keren dat we een steekproef trekken.

Opdrachten voor piraten

1.

a) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

b) De personen in de 'geluk'- en de 'hebzucht'-conditie scoorden gemiddeld het laagst (de twee groepen ontlopen elkaar niet veel). De conditie 'hard werken' deed het beter, en de conditie 'passie' haalde het hoogste gemiddelde. Een verklaring hiervoor kan zijn dat een gepassioneerd rolmodel proefpersonen motiveert om hun best te gaan doen: ze gaan denken dat ze veel geld kunnen verdienen als ze genoeg plezier hebben in hun werk. Ook een rolmodel dat hard werkt kan inspirerend zijn: werk hard en je komt verder. Het is denkbaar dat de proefpersonen in de eerste conditie niet geïnspireerd raakten door een zakenman die vooral veel geluk had gehad, want in dat geval ligt de oorzaak van rijkdom grotendeels buiten je eigen inzet. Ook een hebzuchtig rolmodel zal waarschijnlijk niet inspirerend werken: proefpersonen willen zich met zo iemand niet identificeren.

c) Afhankelijke variabele kwantitatief: voldaan.
Onafhankelijke groepen: voldaan.
Normaliteit: de meeste verdelingen van steekproefscores zijn een beetje scheef naar rechts (bekijk de histogrammen). Echter, de scheefheid en de kurtosis van deze verdelingen zouden in hun respectievelijke populaties best gelijk aan 0 kunnen zijn (check de *Explore*-tabel: alle scheefheid- en kurtosiswaarden liggen tussen -1 en 1). In dat geval komen deze steekproeven toch uit normaal verdeelde populaties. Mocht de normaliteitsassumptie geschonden zijn, dan is dat nauwelijks een probleem: elke steekproef bevat 25 observaties ($n_i = 25$). De ANOVA is dus robuust tegen eventuele schending.
Gelijke varianties: de standaardafwijkingen van de vier steekproeven verschillen van elkaar: de grootste is 69,394 (in de 'hard werken'-conditie) en de kleinste 44,198 (in de 'hebzucht'-conditie). De ratio is $\frac{69{,}394}{44{,}198} = 1{,}57$. Dit verschil valt uiteindelijk dus wel mee: de grootste standaardafwijking is ruim anderhalf keer zo groot

als de kleinste, dus niet twee keer zo groot of nog meer. Daarnaast is *Levene's Test* niet significant ($p = 0{,}115$). We mogen dan ook vermoeden dat er aan deze assumptie is voldaan: het is prima denkbaar dat al deze steekproefstandaardafwijkingen dezelfde σ schatten. Overigens bevat elke steekproef even veel observaties (25), dus de ANOVA zou robuust zijn geweest tegen schending van gelijke varianties.

2. A

MS(Within) is de variantie binnen groepen, en beschrijft de mate waarin individuen binnen één en dezelfde groep nog van elkaar verschillen – oftewel de mate waarin individuen afwijken van hun groepsgemiddelde. De wortel uit de variantie is de standaardafwijking, oftewel de gemiddelde afwijking per persoon. Aangezien *MS(Within)* de variantie binnen álle groepen beschrijft, betreft de grootheid de gehele steekproef. Voor wie dit nog iets verder wil uitpluizen: bedenk dat *MS(Within)* gelijk is aan s_p^2, de gepoolde variantie.
Antwoord B beschrijft de wortel uit *MS(Total)*, een grootheid die normaal niet in de ANOVA-tabel staat omdat we hem niet nodig hebben. $MS(Total) = \frac{SS(Total)}{df(Total)} = \frac{341782{,}990}{99} = 3452{,}35$. De wortel hieruit is de standaardafwijking van alle individuen ten opzichte van het algemene gemiddelde: 58,76. Dat klopt met de *Total Std. Deviation* in de *Descriptives*-tabel! ☺
Antwoord C geeft een grove indruk van wat *MS(Between)* voorstelt.

3. C

Het toetsresultaat is inderdaad heel significant, maar het is niet verstandig de p-waarde te gebruiken om de grootte van het effect *(effect size)* te bepalen; de p-waarde wordt weliswaar beïnvloed door de effectgrootte, maar ook door de omvang van de steekproef (N). Met een grotere steekproef heeft men meer *power*, dus een bijbehorende statistische toets kan een lagere p-waarde hebben – ook als precies hetzelfde effect wordt onderzocht.
De effectmaat η^2 is echter ongevoelig voor de steekproefomvang, en staat voor de proportie verklaarde variatie. $\eta^2 = \frac{SS(Between)}{SS(Total)} = \frac{69814{,}750}{341782{,}990} \approx 0{,}204$. Wat geeft dit aan? Het feit dat niet iedere proefpersoon even veel muntjes verzamelde, kan – op steekproefniveau – voor 20,4% worden verklaard door het feit dat niet iedere proefpersoon hetzelfde rolmodel had. Dat is nogal wat: één enkele factor lijkt al verantwoordelijk voor 0,204 van alle variatie. We kunnen dit dan ook beschouwen als een groot effect. De richtlijnen van Cohen zijn:

- 0,01: klein effect
- 0,06: medium effect
- 0,14: groot effect

Dit zijn natuurlijk maar richtlijnen.

4. B

Aangezien de ANOVA significant is en de nulhypothese kan worden verworpen, is er aantoonbaar een groepseffect. *MS(Between)* wordt daardoor een zuivere schatter van de errorvariantie <u>plus</u> de variantie ten gevolge van dit groepseffect. *MS(Between)* schat dus meer dan *MS(Within)*, zodat de F-ratio een verwachtingswaarde krijgt die groter is dan 1. We weten echter niet precies wat de verwachtingswaarde van F is. We hebben namelijk zowel de errorvariantie als de variantie ten gevolge van het groepseffect moeten schatten. Waarschijnlijk is de F-waarde van deze ANOVA niet precies de F-waarde die we gemiddeld zouden krijgen als we de steekproef van Miyamoto et al. oneindig vaak opnieuw trokken.

5. B

Zowel A als C zegt eigenlijk: 'De kans dat de nulhypothese waar is, is 0,000.' Echter, er bestaat niet zoiets als een kans dat de nulhypothese van Miyamoto et al. waar is[3]; de vier groepen hebben simpelweg hetzelfde populatiegemiddelde of niet. Dat is de 'waarheid'; daar is geen kans mee gemoeid. Het idee van een statistische toets is dat we uitgaan van de nulhypothese: we zijn sceptisch en nemen aan dat de groepen qua gemiddelde <u>niet</u> van elkaar verschillen. Vervolgens kijken we of de getrokken steekproef zó bijzonder is dat we dit uitgangspunt kunnen verwerpen. Het blijkt dat deze steekproef, met zulke verschillen tussen de groepen, vrijwel nooit zou voorkomen (minder dan 1 op de 1000 keer) als de vier populatiegemiddelden hetzelfde waren. Dat is genoeg reden om niet langer te geloven dat de vier populatiegemiddelden gelijk zijn.

6. A

[3] Althans: niet in de frequentistische opvatting van het begrip 'kans'. Wel in bayesiaanse statistiek, mocht je het interessant vinden…

Let op: we hebben vier groepen, dus er ontstaan zes unieke vergelijkingen! Tel het aantal rijen in de tabel maar eens (12), en deel dit door 2. Of gebruik het rekenregeltje $\frac{k(k-1)}{2}$, waarbij k staat voor het aantal groepen: $\frac{4*(4-1)}{2} = \frac{4*3}{2} = 6$. Als we alle p-waarden verzesvoudigen, krijgen we de volgende tabel:

Multiple Comparisons

Dependent Variable: muntjes

Bonferroni

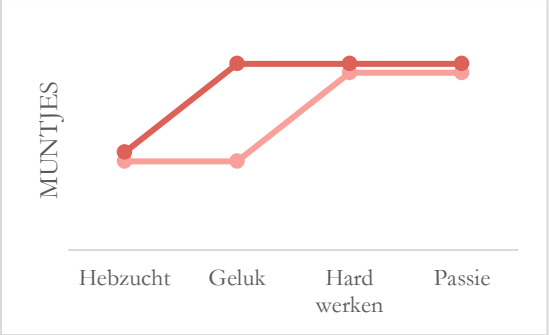
(I) rolmodel	(J) rolmodel	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
geluk	hard werken	-38,600	15,055	,071	-79,16	1,96
	passie	-59,400*	15,055	,001	-99,96	-18,84
	hebzucht	3,400	15,055	1,000	-37,16	43,96
hard werken	geluk	38,600	15,055	,071	-1,96	79,16
	passie	-20,800	15,055	1,000	-61,36	19,76
	hebzucht	42,000*	15,055	,038	1,44	82,56
passie	geluk	59,400*	15,055	,001	18,84	99,96
	hard werken	20,800	15,055	1,000	-19,76	61,36
	hebzucht	62,800*	15,055	,000	22,24	103,36
hebzucht	geluk	-3,400	15,055	1,000	-43,96	37,16
	hard werken	-42,000*	15,055	,038	-82,56	-1,44
	passie	-62,800*	15,055	,000	-103,36	-22,24

*. The mean difference is significant at the 0.05 level.

(Opmerking: een p-waarde kan niet groter worden dan 1 ofwel 100%.)
Na deze toch wel zware correctie zijn er nog slechts drie unieke vergelijkingen significant:

- Het verschil tussen 'geluk' en 'passie';
- Het verschil tussen 'hard werken' en 'hebzucht';
- Het verschil tussen 'hebzucht' en 'passie'.

In dit resultaat zit iets tegenstrijdigs, want probeer het maar eens te plotten:

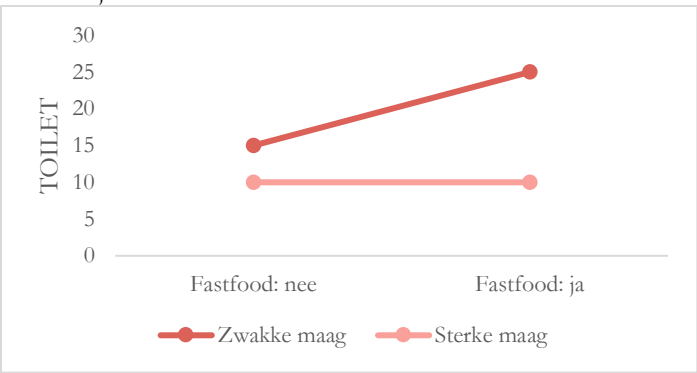


'Geluk' verschilt aantoonbaar van 'passie': in dat geval zou het lichte patroon de populatiegemiddelden het best weergeven. Maar 'geluk' verschilt niet langer aantoonbaar van 'hard werken': in dat geval zou het donkere patroon correct zijn. De twee patronen zijn onverenigbaar en dus moet er ergens sprake zijn van een Type I- of Type II-fout. Omdat 'geluk' en 'hard werken' vóór de Bonferroni-correctie nog wel significant van elkaar verschilden, maken we nu waarschijnlijk bij deze vergelijking een Type II-fout.

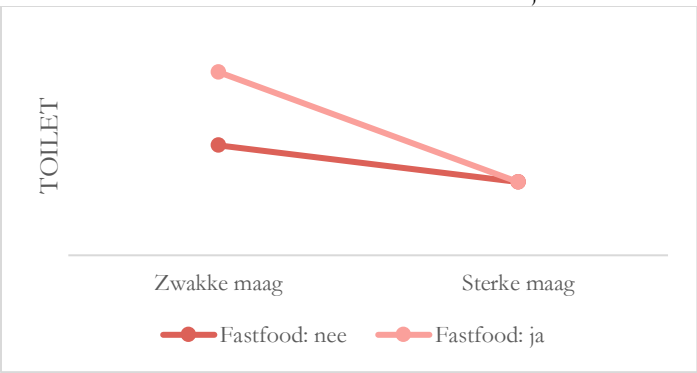
HOOFDSTUK 11 TWEEWEG-ANOVA

1.

a) Persoonlijk dacht ik aan een interactie-effect:



Buikgrieppatiënten met een zwakke maag krijgen misschien wel extra veel last van diarree als ze fastfood hebben gegeten. Voor patiënten met een sterke maag maakt het eten van fastfood misschien wel niet uit.
We kunnen het interactie-effect ook andersom bekijken:



Patiënten met een zwakke maag moeten gemiddeld vaker naar het toilet dan patiënten met een sterke maag, maar het verschil is groter als de patiënten fastfood hebben gegeten.
Nogmaals: andere ideeën, mits goed onderbouwd, zijn ook goed.

b) Dit is een non-orthogonaal design, hoewel het niet veel scheelt. De proefpersonen zijn namelijk niet helemaal in een constante verhouding over de condities verdeeld:

	Zwakke maag	Sterke maag
Fastfood: nee	19	20
Fastfood: ja	16	16

Als de verhoudingen allebei 19/16 of allebei 20/16 waren geweest, waren de factoren fastfood en maag statistisch onafhankelijk geweest. Nu vertonen ze een lichte samenhang en bestaat het risico van milde *confounding*. We zien dit in de *Descriptive Statistics*-tabel.
Om te corrigeren voor eventuele *confounding*, laten we SPSS in de ANOVA *Type III Sums of Squares* berekenen. In een orthogonaal design hoeven de *Sums of Squares* van de losse effecten niet te worden gecorrigeerd en tellen deze precies op tot het *Corrected Model*, maar hier niet: $623{,}754 + 140{,}695 + 323{,}719 = 1088{,}168 \neq 1010{,}406$.
Tevens kunnen we SPSS vragen om *Estimated Marginal Means*. Deze groepsgemiddelden zijn ongewogen, en daardoor gecorrigeerd voor eventuele *confounding*. Ze verschillen dan ook lichtjes van de gemiddelden in de *Descriptive Statistics*-tabel.

c) Afhankelijke variabele kwantitatief: voldaan.
Onafhankelijke groepen: voldaan.
Normaliteit: de meeste groepen zijn duidelijk scheef verdeeld: in elke conditie waren er wel wat buikgrieppatiënten die zo'n halve dag langer op het toilet hebben doorgebracht dan de meesten. Diverse normaliteitstoetsen zijn significant. We mogen dus niet aannemen dat er aan de assumptie van normaliteit is voldaan.
De vraag is nu of de ANOVA robuust is tegen schending van deze assumptie. Twee condities tellen slechts 16 proefpersonen. Dit is weinig. Volgens de nieuwste editie van Moore, McGabe & Craig (zie de bronvermelding in het handboek en elders op de website) is het voldoende als de steekproef minimaal 15 observaties kent, én niet extreem scheef is, én geen uitschieters bevat. Het is dan voor discussie vatbaar wat

je als extreem scheef moet beschouwen. Op zich zou ik wel durven door te gaan met de ANOVA. Ik zou echter alleen toetsresultaten vertrouwen die duidelijk heel significant waren, of duidelijk helemaal niet significant.
Gelijke varianties: aan deze assumptie is duidelijk niet voldaan. De grootste en de kleinste standaardafwijking verschillen fors van elkaar $\left(\frac{9,571}{2,903} = 3,30 > 2\right)$ en *Levene's Test* is zeer significant ($p = 0,001$). Is de ANOVA robuust tegen deze schending? Dat valt te betwijfelen: de groepen zijn niet even groot. Jouw conclusie, beste lezer, mag luiden: 'Nee, de ANOVA is niet robuust en we mogen hem eigenlijk niet uitvoeren.'[4]

2. B

Dit design is non-orthogonaal, dus zolang de interactieterm in het model zit, vallen de hoofdeffecten niet te interpreteren. A spreekt over een hoofdeffect van de maag en is daarom fout. B definieert het interactie-effect en dit is inderdaad significant ($p = 0,007$). C ten slotte heeft het over een samenhang tussen de twee onafhankelijke variabelen, maag en fastfood (dus niet over een interactie-effect!). Deze samenhang is er enigszins in de steekproef: de mensen met een zwakke maag hebben iets vaker $\left(\frac{16}{35}\right)$ fastfood gegeten dan de mensen met een sterke maag $\left(\frac{16}{36}\right)$ (oftewel, het design is non-orthogonaal). Het verschil is echter minuscuul en zal er in de populatie waarschijnlijk niet zijn. Sowieso wordt het nergens in de uitvoer getoetst!

3. B

Een niet-significante factor kan in het algemeen beter worden verwijderd, want er moet rekening mee worden gehouden en dat kost *power*. Echter, het effect van de maag is al zeer significant en kan alleen maar significanter worden als de factor fastfood wordt verwijderd; de errorvariantie zal dan namelijk dalen.

4. A

In de steekproef zien we wel een effect: de groepen die fastfood hebben gegeten, moeten gemiddeld wat vaker naar het toilet. Dat uit zich ook in de *Sum of Squares* van de factor fastfood: als fastfood ook in de steekproef geen enkel effect had, zou deze *Sum of Squares* gelijk zijn aan nul. Echter, het steekproefeffect is te klein om aan te tonen dat er ook in de populatie een effect is ($p = 0,089$); vooralsnog moeten we het steekproefeffect aan toeval toeschrijven.
Opmerking: in een steekproef doet zich eigenlijk bijna altijd een effect voor, omdat het nu eenmaal een steekproef is en groepen door toeval zullen verschillen. De statistische toets dient om na te gaan of de verschillen komen door meer dan enkel toeval.

5. B

Antwoord A is complete onzin; hoe moeten we ons dat voorstellen? Als we zeggen dat een zwakke maag een effect heeft, bedoelen we dat altijd in vergelijking tot iets anders. Mensen met een zwakke maag moeten vaker naar het toilet dan… mensen met een sterke maag natuurlijk. In dat geval heeft een sterke maag automatisch ook een effect: je moet minder vaak naar het toilet dan… met een zwakke maag. Kijken we alleen naar de mensen met een zwakke maag, dan is de maag geen variabele meer (iedereen scoort er tenslotte hetzelfde op) en zoals ik hierboven aantoon, kunnen we alleen het effect toetsen van variabelen. Welke variabele zal dan de factor zijn in deze ANOVA's? Dat moet dan wel fastfood zijn.
Bij mensen met een zwakke maag wordt er een effect van fastfood aangetoond ($p = 0,012$); bij mensen met een sterke maag niet ($p = 0,345$). Dit is consistent met het interactie-effect dat we eerder vonden. De nog nodige Bonferroni-correctie (beide p-waarden maal 2) verandert de conclusies niet.

6. C

Nogmaals: kijk in het volledige model eerst naar de interactieterm. Is deze niet significant, dan dienen we hem te verwijderen om de hoofdeffecten te kunnen interpreteren (bij een orthogonaal design mag hij ook blijven staan). Is de interactieterm wel significant, zoals nu, dan hebben hoofdeffecten (vaak) geen betekenis meer. We moeten dan simpele effecten gaan onderzoeken.

[4] Je zou nog kunnen verdedigen dat de twee meest extreme standaardafwijkingen ook uit de kleinste steekproeven komen (en dus minder zwaar meewegen), en de twee meer gematigde standaardafwijkingen uit de grootste steekproeven, maar dat voert voor een normale statistiekcursus wat te ver.

HOOFDSTUK 12 ANCOVA

1. B

Het is niet waarschijnlijk dat 'emotionaliteit vooraf' een *confounder* is; de proefpersonen zijn willekeurig toegewezen aan een 'muziek'-conditie en dat maakt de kans op samenhang tussen 'muziek' en 'emotionaliteit vooraf' klein. Op zich zouden hoog emotionele mensen anders kunnen reageren op muziek dan laag emotionele mensen (interactie), maar een meer voor de hand liggende verklaring voor de niet-significante ANOVA is dat de errorvariantie gewoon te groot is: de deelnemers verschillen qua emotionaliteit weliswaar van elkaar door de muziek die ze luisteren, maar ook door allerlei andere dingen, waardoor dit muziekeffect niet kan worden gedetecteerd. Er is te veel ruis.

2. A

Hier wordt de nulhypothese getoetst dat de drie groepen hetzelfde populatiegemiddelde hebben op 'emotionaliteit vooraf'; deze nulhypothese kan totaal niet worden verworpen. A is dan ook een correcte uitspraak. Het betekent dat 'emotionaliteit vooraf' definitief geen *confounder* is, maar er is nog een andere reden denkbaar om de variabele als covariaat op te nemen: hij verkleint waarschijnlijk de errorvariantie aanzienlijk. Een groot deel van de ruis waar we in opgave 1 tegenaan liepen, kunnen we wegnemen door te kijken wie er vooraf emotioneel was en wie niet.

3. A

De normaliteitsassumptie kan niet worden gecontroleerd. De assumptie van gelijke varianties is inderdaad geschonden (*Levene's Test*: $p = 0{,}003$), maar de drie groepen zijn even groot. Wat de aanvullende assumpties van ANCOVA betreft: in het spreidingsdiagram zien we geen kromlijnige verbanden, dus aan lineariteit is voldaan. Verder is de interactietoets in Model 4 niet significant, dus ook aan non-interactie is voldaan.

4. B

Het spreidingsdiagram in Model 3 toont het steekproefresultaat: in de steekproef lopen de drie regressielijnen niet helemaal parallel en dus zijn de b's niet helemaal gelijk. Kortom, er doet zich enige interactie voor. Echter, deze interactie is niet significant (zie Model 4).

5. C

Model 1 hadden we al eerder afgeschreven omdat het niet genoeg *power* heeft (te veel errorvariantie). Model 4 bevat nog een interactieterm en is dus een SCHMANCOVA; het hoofdeffect van 'muziek' mag hier niet geïnterpreteerd worden. Het effect van 'muziek' wordt valide getoetst in Model 5. Het blijkt significant ($p = 0{,}038$).

6. A

Deze vraag gaat over de manier waarop ANCOVA corrigeert voor eventuele *confounding* (zie de uitleg in de theorie). Maak hiervan een tekening. Het gekozen referentiepunt is momenteel een emotionaliteit vooraf gelijk aan 51,49 (dit staat onder de *EMM*-tabelletjes). Dit is de gemiddelde emotionaliteit vooraf in de complete steekproef. Als iedereen vooraf 10 punten minder emotioneel was geweest, zou het referentiepunt op 41,49 zijn komen te liggen. Voor iemand die vooraf minder emotioneel was, voorspelt het ANCOVA-model ook na de computertaak een lagere emotionaliteit (kijk maar naar de regressielijn). Alle groepen krijgen dus een lager *Estimated Marginal Mean*. Omdat ANCOVA uitgaat van non-interactie, lopen de lijnen echter wel parallel: de drie *Estimated Marginal Means* dalen daarom precies even ver. Hun onderlinge verschillen zullen dus gelijk blijven.

HOOFDSTUK 13 *WITHIN-SUBJECTS*-ANOVA

1. A

'Proefpersoon' is een noodzakelijke *random factor* voor een *within-subjects*-design, zelfs als deze geen aantoonbaar effect heeft op de afhankelijke variabele; dat zal in de populatie namelijk heus wel zo zijn. Als we 'proefpersoon' als *random factor* toevoegen, wordt het een *within-subjects*-design; SPSS ziet dat niet en denkt dat 'proefpersoon' een tweede *between-subjects*-factor is. Dit maakt verder niet uit voor de analyse. Het enige serieuze probleem van *GLM Univariate* wordt beschreven in antwoord A: wordt er niet aan sfericiteit voldaan, dan heb je een epsiloncorrectie nodig. *GLM Univariate* voert deze niet uit.

2. B

Verder achter de komma verschillen de p-waarden van de *Multivariate Tests* en de *Tests of Within-Subjects Effects* wel een beetje; de toetsen zijn niet exact gelijk. Wel toetsen ze dezelfde nulhypothese.

3. A

De p-waarde van *Mauchly's Test* is niet de juiste: deze toets gaat na of er sfericiteit is (nulhypothese: ja) en is significant. Willen we dan ook het effect van het aantal verkeersdoden op het dodental toetsen, dan kunnen we niet de gewone univariate toets gebruiken; een epsiloncorrectie is nodig, bij voorkeur *Greenhouse-Geisser*. 0,694 vormt de epsilonschatting (zie de theorie voor details), niet de p-waarde van de *Greenhouse-Geisser*-toets. Kijk in de *Tests of Within-Subjects Effects* in de rij *Greenhouse-Geisser*. De p-waarde die hier staat is wat we zoeken.

4. B

Lower-bound is een zware epsiloncorrectie, maar dat betekent niet dat de bijbehorende toets nooit significant kan worden; hij heeft relatief weinig *power*, maar natuurlijk niet helemaal geen. De andere toetsen zijn extreem significant, hetgeen suggereert dat *Lower-bound* dat ook wel zal zijn. Helemaal zeker kun je het echter niet weten.

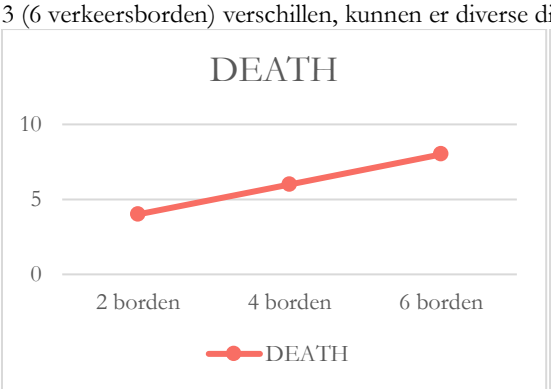
5. C

In *GLM Repeated Measures* heeft elke paarsgewijze vergelijking netjes zijn eigen standaardfout, omdat dit programma niet uitgaat van sfericiteit. Ga je wel uit van sfericiteit, zoals *GLM Univariate*, dan beschouw je die verschillende standaardfouten als schatters van dezelfde. Je berekent dan een gewogen gemiddelde: sommige standaardfouten worden groter, sommige kleiner. Dientengevolge worden ook de corresponderende p-waarden groter respectievelijk kleiner.

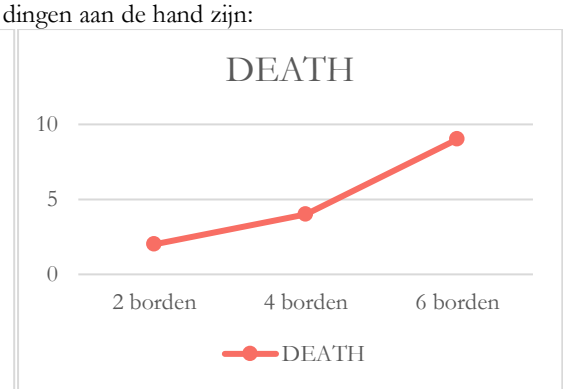
6.

a) Onderzoekster Clarkson verwachtte een <u>trend</u>: een kwadratische trend welteverstaan. Het aantal aangereden fietsers en voetgangers neemt kwadratisch toe met het aantal verkeersborden langs de weg. Met de *Tests of Within-Subjects Effects* kunnen we alleen vaststellen of er een algemeen effect is van het aantal verkeersborden, niet hoe dit effect eruitziet: de nulhypothese is dat alle drie de condities hetzelfde populatiegemiddelde hebben, de alternatieve hypothese is dat minstens één gemiddelde verschilt. Als het aantal aanrijdingen steeds sneller toeneemt bij toename van het aantal verkeersborden, moeten alle populatiegemiddelden verschillen. Dit laatste kunnen we met de univariate ANOVA niet aantonen.

b) De paarsgewijze vergelijkingen kunnen ons wel vertellen welke gemiddelden van elkaar verschillen. Echter, we kunnen hooguit twee condities tegelijk vergelijken. Daarmee zegt ook deze analyse niet zoveel over de precieze verwachting van onderzoekster Clarkson. Als bijvoorbeeld conditie 2 (4 verkeersborden) en conditie 3 (6 verkeersborden) verschillen, kunnen er diverse dingen aan de hand zijn:



Lineair



Kwadratisch

Daarom zijn ook de paarsgewijze vergelijkingen niet specifiek genoeg. Daarbovenop komt nog eens dat ze minder *power* hebben vanwege een noodzakelijke Bonferroni-correctie.

c) De kwadratische trend die Clarkson verwachtte, kan wél rechtstreeks worden onderzocht in de *Tests of Within-Subjects Contrasts*-tabel en wordt aangetoond ($p = 0{,}008$). Dit contrast is van een hogere orde dan het eveneens significante lineaire contrast en gaat dus in principe voor (zie de theorie).

d) Is het inhoudelijk zinnig om een contrastanalyse te gebruiken, kies deze dan altijd. De categorieën van de onafhankelijke variabele (borden) vormen eigenlijk een kwantitatieve schaal. Daarom raad ik contrastanalysen aan.

HOOFDSTUK 14-15 TWEEWEG-*WITHIN-SUBJECTS*- EN *SPLIT-PLOT*-ANOVA

1.

a) De rondetijd.
b) Beenlengte: *between-subjects*-factor (lengte verschilt tussen personen). Trap: *within-subjects*-factor (alle personen proberen beide trappen). Dit betekent een *split-plot*-design.
c) *Meerdere antwoorden mogelijk*. Maak een schets door de afhankelijke variabele op de Y-as te zetten, en de *within-subjects*-factor op de X-as. Trek nu een lijntje per *between-subjects*-groep.

2. B

De assumptie van gelijke covariantiematrices stelt dat de *between-subjects*-groepen:

- Dezelfde variantie hebben in rondetijd op de luie trap (de eerste *Levene's Test* controleert hiervoor);
- Dezelfde variantie hebben in rondetijd op de normale trap (de tweede *Levene's Test* controleert hiervoor);
- Dezelfde covariantie (correlatie) hebben tussen hun rondetijden op de luie en de normale trap (hier controleert geen van beide *Levene's Tests* voor).

Om te toetsen of de gehele covariantiematrices gelijk zijn, hebben we *Box's Test* nodig. In tegenstelling tot *Box's Test* is *Levene's Test* robuust tegen schending van normaliteit als de steekproeven maar groot genoeg zijn.

3. C

Het feit dat de factor beenlengte 3 niveaus heeft, maakt niets uit: dit is de *between-subjects*-factor! De multi- en univariate toets komen exact overeen als de *within-subjects*-factor 2 niveaus heeft, en dit is gewoon het geval.

4. C

Kijk eerst naar het interactie-effect: significant! Hoofdeffecten zoals beschreven in A en B zijn nu niet zinvol meer. De simpele effecten van de trap verschillen: mensen met korte benen zijn het snelst op een luie trap, maar mensen met gemiddelde en lange benen doen het beter op een normale trap. Conclusie: de ideale trap bestaat niet.

5. A

Antwoord B beschrijft slechts het steekproefresultaat; dat kan ook toevallig zijn en betekent niet dat er ook in de populatie een effect van beenlengte is. Interessanter is dat zich inderdaad een significant interactie-effect voordoet. We mogen nu niet meer naar hoofdeffecten kijken… maar de paarsgewijze vergelijkingen toetsen het <u>hoofdeffect</u> van beenlengte! De trap die de deelnemers afleggen, speelt bij deze vergelijkingen immers geen rol. De paarsgewijze vergelijkingen zijn betekenisloos: aangezien er interactie is, hangt het beenlengte-effect af van de trap.

6. B

Aangezien we interactie hebben gevonden, moet er nog een analyse van simpele effecten achteraan. C gaat over een hoofdeffect van de trap. Paarsgewijze vergelijkingen zijn daartoe sowieso niet nodig, want trap heeft maar 2 niveaus; de ANOVA is al voldoende.

HOOFDSTUK 16-19 CATEGORISCHE TOETSEN

1.

a) We bekijken nu twee categorische variabelen: de plaats waar de strandgangers vooral hun tijd hebben doorgebracht, en of ze schoenen hebben gedragen. Een χ^2-kruistabeltoets is dan sowieso geschikt. Beide variabelen zijn dichotoom, dus de kruistabel heeft $(2-1)*(2-1)=1$ vrijheidsgraad. Daarom is ook een z-toets voor 2 proporties een optie.

b) **Z-toets voor 2 proporties**

$$H_0: \pi_{strand} = \pi_{paviljoen}$$
$$H_A: \pi_{strand} \neq \pi_{paviljoen}$$

π is hier de proportie personen die op blote voeten lopen.
Aan de designassumpties is voldaan (afhankelijke variabele dichotoom, onafhankelijke groepen). Normaliteit is sowieso geschonden, maar de steekproeven zijn qua grootte ruim voldoende.
Dus uitvoeren maar:

$$p_1 = \frac{121}{187} = 0{,}647$$
$$p_2 = \frac{22}{48} = 0{,}458$$
$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{121+22}{187+48} = 0{,}609$$
$$Z = \frac{p_1 - p_2}{\sqrt{\pi(1-\pi)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{0{,}647 - 0{,}458}{\sqrt{0{,}609 * 0{,}391} * \sqrt{\frac{1}{187}+\frac{1}{48}}} = \frac{0{,}189}{0{,}488 * 0{,}162} = 2{,}39$$

De z-tabel in de appendix vertelt ons dat

$$p = 2 * P(Z > 2{,}39) = 2 * (1 - 0{,}9916) = 2 * 0{,}0084 = 0{,}0168$$

Vergeet de p-waarde niet te verdubbelen indien je zoals ik een tweezijdige toets uitvoert.
Enfin, de uitkomst is significant: mensen op het strand lopen aantoonbaar vaker op blote voeten.

χ^2-kruistabeltoets

$$H_0: geen\ samenhang\ plaats \times schoeisel$$
$$H_A: wel\ samenhang\ plaats \times schoeisel$$

Aan de designassumpties is voldaan (afhankelijke variabele categorisch, onafhankelijke groepen). Of de *Expected Counts* groot genoeg zijn zullen we zo merken, maar gezien de grote steekproeven is dat alvast erg aannemelijk.
Dus uitvoeren maar. Eerst maken we de kruistabel:

Observed		Waar de dag doorgebracht?		
		op het strand	in het paviljoen	
Hoe rondgelopen?	op blote voeten	121	22	143
	in schoenen	66	26	92
		187	48	235

De *Expected Counts* berekenen we met de formule $EC = \frac{rijtotaal * kolomtotaal}{N}$. Ze blijken inderdaad allemaal groter dan 5:

Expected		Waar de dag doorgebracht?		
		op het strand	in het paviljoen	
Hoe rondgelopen?	op blote voeten	113,8	29,2	143
	in schoenen	73,2	18,8	92
		187	48	235

Dan nu de toetsingsgrootheid:

$$\chi^2 = \sum \frac{(OC - EC)^2}{EC} =$$
$$\frac{(121-113{,}8)^2}{113{,}8} + \frac{(66-73{,}2)^2}{73{,}2} + \frac{(22-29{,}2)^2}{29{,}2} + \frac{(26-18{,}8)^2}{18{,}8} =$$
$$0{,}456 + 0{,}708 + 1{,}775 + 2{,}757 = 5{,}696$$

Ga maar na: $Z^2 = \chi^2$. Jazeker: de z-toets en de χ^2-toets zijn in dit geval precies hetzelfde – oftewel, met een moeilijk woord, data-equivalent. We kunnen ze dus allebei gebruiken, indien de kruistabel slechts 1 vrijheidsgraad heeft! Je kunt meer over de gelijkenis lezen in brughoofdstuk 29 van het handboek.
Hoe dan ook, de χ^2-tabel in de appendix vertelt ons (kijk bij 1 vrijheidsgraad) dat

$$5{,}412 < \chi^2 < 6{,}635$$

$$0{,}01 < p < 0{,}02$$

De uitkomst is dus in elk geval significant: mensen op het strand lopen aantoonbaar vaker op blote voeten. Dat lijkt mij ook netter; in een paviljoen trek je wat aan, of niet?

c) Dat gaat alleen maar zomaar indien je voor een z-toets gekozen hebt!
Het betrouwbaarheidsinterval is

$$p_1 - p_2 \pm Z^* \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

$$0{,}647 - 0{,}458 \pm 1{,}96 * \sqrt{\frac{0{,}647 * 0{,}353}{187} + \frac{0{,}458 * 0{,}542}{48}}$$

$$0{,}189 \pm 1{,}96 * \sqrt{0{,}00122 + 0{,}00517}$$

$$0{,}189 \pm 0{,}157$$

$$[\,0{,}032\,;\,0{,}346]$$

Het verschil in proportie blotevoetenlopers bedraagt in de populatie dus waarschijnlijk tussen de 3,2 en 34,6 procent.

2.
We hebben één variabele met meer dan 2 categorieën. De tabel heeft $3 - 1 = 2$ vrijheidsgraden; een χ^2-*Goodness of Fit*-toets is dus het enige alternatief.
In totaal zijn er 235 proefpersonen ($N = 235$). We berekenen de *Expected Counts* met de formule $EC_i = N * \pi_i$:

$$EC_{ja} = 235 * 0{,}3 = 70{,}5$$

$$EC_{een\ beetje} = 235 * 0{,}3 = 70{,}5$$

$$EC_{nee} = 235 * 0{,}4 = 94$$

Deze zijn allemaal ruim voldoende.
Nu kunnen we de toetsingsgrootheid berekenen:

$$\chi^2 = \sum \frac{(OC - EC)^2}{EC} = \frac{(45 - 70{,}5)^2}{70{,}5} + \frac{(66 - 70{,}5)^2}{70{,}5} + \frac{(124 - 94)^2}{94} =$$

$$9{,}22 + 0{,}29 + 9{,}57 = 19{,}08$$

Deze waarde is bij 2 vrijheidsgraden zwaar significant: groter dan de grootste waarde in de appendixtabel, dus de p-waarde is nog kleiner dan 0,0005. In Moonshine houden de strandgangers duidelijk vaker van zand.

3.

a) Aangezien alle *Observed Counts* veel groter zijn dan 5, zullen ook alle *Expected Counts* dat zijn (we kunnen dus een stukje afsnijden). Voor het toetsresultaat kijken we in de *Chi-Square Tests*-tabel: de *Pearson Chi-Square* heeft een p-waarde van 0,008 en is dus overtuigend significant. Afhankelijk van hun hekel aan zand, lopen de strandgangers van Moonshine niet even vaak op blote voeten.

b) Ja dus: <u>paarsgewijze vergelijkingen</u>! Met Bonferroni-correctie. Het rare is dat deze voor kruistabellen niet in SPSS zitten. Daarom doet niemand ze, hoewel het natuurlijk een goed gebruik zou zijn…
Als voorvechter van degelijke statistiek (je moet niet achter de massa aanlopen!), doe ik de paarsgewijze vergelijkingen nu toch:

Hoe rondgelopen? * Hekel aan zand? Crosstabulation

Count

		Hekel aan zand?		Total
		een beetje	nee	
Hoe rondgelopen?	in schoenen	32	37	69
	op blote voeten	34	87	121
Total		66	124	190

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6,475[a]	1	,011		
Continuity Correction[b]	5,694	1	,017		
Likelihood Ratio	6,392	1	,011		
Fisher's Exact Test				,017	,009
Linear-by-Linear Association	6,441	1	,011		
N of Valid Cases	190				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 23,97.

b. Computed only for a 2x2 table

Na Bonferroni-correctie[5] geldt $p = 0{,}033$: significant. Mensen met een beetje hekel aan zand lopen minder vaak op blote voeten dan mensen zonder hekel.

Hoe rondgelopen? * Hekel aan zand? Crosstabulation

Count

		Hekel aan zand?		Total
		ja	nee	
Hoe rondgelopen?	in schoenen	23	37	60
	op blote voeten	22	87	109
Total		45	124	169

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6,525[a]	1	,011		
Continuity Correction[b]	5,629	1	,018		
Likelihood Ratio	6,356	1	,012		
Fisher's Exact Test				,017	,009
Linear-by-Linear Association	6,486	1	,011		
N of Valid Cases	169				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 15,98.

b. Computed only for a 2x2 table

Na Bonferroni-correctie krijgen we $p = 0{,}033$: significant. Mensen met een uitgesproken hekel aan zand lopen minder vaak op blote voeten dan mensen zonder hekel.

Hoe rondgelopen? * Hekel aan zand? Crosstabulation

Count

		Hekel aan zand?		Total
		ja	een beetje	
Hoe rondgelopen?	in schoenen	23	32	55
	op blote voeten	22	34	56
Total		45	66	111

[5] Ik verdriedubbel hier de p-waarde, maar je mag ook het significantieniveau α door 3 delen. Zie hoofdstuk 10.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,074[a]	1	,786		
Continuity Correction[b]	,006	1	,938		
Likelihood Ratio	,074	1	,786		
Fisher's Exact Test				,848	,469
Linear-by-Linear Association	,073	1	,787		
N of Valid Cases	111				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 22,30.

b. Computed only for a 2x2 table

Na Bonferroni-correctie heb je $p = 1$ (een p-waarde kan niet groter worden dan 100%). Mensen met een duidelijke en een beetje hekel aan zand lopen dus waarschijnlijk even vaak op blote voeten.

4.
Het aantal gekochte ijsjes is kwantitatief. Hier is een χ^2-toets ongeschikt voor. Ze heeft een *independent samples t-test* nodig of een eenweg-ANOVA.

5. C
Er zijn weliswaar meer debutanten die sluikreclame maken, maar de totale steekproef telt ook veel meer debutanten: het is lekker eerlijk vergelijken zo! A is dus fout. B komt al meer in de buurt: als de relatieve frequentieverdelingen van elkaar verschillen, betekent dat dat de proportie sluikreclamemakers bij de debutanten ongelijk is aan die bij de gevestigde auteurs. Dit is zo (25% sluikreclamemakers bij de debutanten tegenover slechts 12,5% bij de gevestigde auteurs), dus de twee verdelingen verschillen… op steekproefniveau. De vraag is of dit verschil niet gewoon toevallig plaatsvindt, en of deze steekproeven getrokken zijn uit twee gelijke populatieverdelingen waarin even veel sluikreclame gemaakt wordt. Daarom is ook B onjuist. Dat maakt C automatisch correct. Wie bevestiging wil, kan doorgaan.
In dat geval toetsen we de nulhypothese – uiteraard met een χ^2-toets. Eerst berekenen we de *Expected Counts*.

	Debutant	Gevestigde auteur	
Maakt geen sluikreclame	45 (48)	35 (32)	**80**
Maakt wel sluikreclame	15 (12)	5 (8)	**20**
	60	**40**	**100**

Vervolgens: $\boldsymbol{\chi^2} = \frac{(45-48)^2}{48} + \frac{(15-12)^2}{12} + \frac{(35-32)^2}{32} + \frac{(5-8)^2}{8} = \mathbf{2,34}$
Uit de χ^2-tabel blijkt dat bij 1 vrijheidsgraad $0{,}10 < P(\chi^2 > 2{,}34) < 0{,}15$. De p-waarde is dus sowieso groter dan 0,05.[6]
De onderzoeker verwerpt de nulhypothese dus niet, maar het zou kunnen dat hij een Type II-fout maakt.
Dit had ook gekund met een z-toets voor 2 proporties. Het voordeel daarvan is dat de onderzoeker een eenzijdige toets had kunnen uitvoeren. Deze zou echter nog steeds niet significant zijn geweest; de p-waarde (die half zo groot is) blijft te groot.

6. A
De *Expected Counts* van die derde schrijverscategorie zijn veel te klein (ze moeten minimaal 5 zijn). Probeer ze maar eens uit te rekenen!

[6] Overigens vertelt SPSS ons: $p = 0{,}126$.

HOOFDSTUK 20 KRUISTABELANALYSE

1.

a)

Codeer steeds 'niet' als 0, en 'wel' als 1. In dat geval moet 'niet' steeds links of boven, en 'wel' steeds rechts of onder.

	lage SES		hoge SES	
	geen keukenpapier	keukenpapier	geen keukenpapier	keukenpapier
ongelukkig	49	32	40	26
gelukkig	96	51	91	43
	145	85	131	69

b) Met deze fatsoenlijk gestructureerde kruistabel kunnen we *odds*-ratio's uitrekenen zonder bang te zijn voor fouten. Voor de veiligheid (om *confounding* te vermijden) bekijken we de simpele effecten van KEUKENPAPIER.

- Lage SES: $OR_{GELUKKIG*KEUKENPAPIER} = \frac{A*D}{B*C} = \frac{49*51}{32*96} = 0{,}81$
- Hoge SES: $OR_{GELUKKIG*KEUKENPAPIER} = \frac{40*43}{26*91} = 0{,}73$

Al met al zien we bij beide groepen een effect: de *odds*-ratio is kleiner dan 1, dus er is een negatief verband. Het gebruik van keukenpapier verlaagt je *odds* om gelukkig te zijn. Let wel: dit effect zou door steekproeftoeval kunnen komen.

c) Voor de veiligheid (om *confounding* te vermijden) bekijken we de simpele effecten van SES. Let er goed op dat je de juiste cellen uit de kruistabel plukt. Indien nodig: dek de kolommen die niet meedoen even af!

- Geen keukenpapier: $OR_{GELUKKIG*SES} = \frac{A*D}{B*C} = \frac{49*91}{40*96} = 1{,}16$
- Keukenpapier: $OR_{GELUKKIG*SES} = \frac{32*43}{26*51} = 1{,}04$

Al met al zien we bij beide groepen een effect: de *odds*-ratio is groter dan 1, dus er is een positief verband. Mensen met een hogere SES zijn iets vaker gelukkig. Let wel: dit effect zou door steekproeftoeval kunnen komen.

d) Het feit dat de simpele effecten (de *odds*-ratio's) steeds verschillen, betekent dat er interactie is: het effect van KEUKENPAPIER op GELUKKIG hangt deels af van SES. Automatisch hangt het effect van SES op GELUKKIG dan ook deels af van KEUKENPAPIER (interactie is symmetrisch, noemen we dat).

2. C

Antwoord A stelt dat het interactie-effect in de steekproef gering is. Dat is misschien wel zo, maar interactie heeft niets te maken met *confounding*!

Of het gecorrigeerde hoofdeffect van KEUKENPAPIER groot of klein is, valt wel enigszins te betwisten. Ik vind het niet zo groot, beste lezer. Maar die discussie maakt ook niets uit: als we niet hadden gecorrigeerd voor *confounding*, hadden we misschien wel een veel groter effect of zelfs helemaal geen effect gevonden van KEUKENPAPIER! Het hele punt is of de correctie voor *confounding* je resultaten drastisch verandert of niet. Daarmee valt ook antwoord B af.

Er is maar één manier om het risico van *confounding* in kaart te brengen: maak een aparte kruistabel van de twee onafhankelijke variabelen. Negeer de afhankelijke variabele GELUKKIG.

		SES	
		laag	hoog
KEUKENPAPIER	nee	145	131
	ja	85	69

De *odds*-ratio blijkt gelijk aan

$$OR_{KEUKENPAPIER*SES} = \frac{145*69}{131*85} = 0{,}90$$

Het – negatieve – verband is dus niet al te sterk (1 betekent geen verband). Mensen met een lage SES gebruiken in deze steekproef iets vaker keukenpapier, maar het houdt niet over. Eventuele *confounding* zal dus gering zijn. Het juiste antwoord is C.

3. A

Antwoord A en B verwijzen allebei naar de *Mantel-Haenszel Common Odds Ratio Estimate*. De vraag is dus: welk verband drukt deze *odds*-ratio precies uit? SES is de covariaat en fungeert als controlevariabele; hij wordt diverse keren genoemd aan de linkerkant van tabellen. Het effect van de controlevariabele wordt niet getoetst! We moeten dus wel de *odds*-ratio zien voor het verband tussen GELUKKIG en KEUKENPAPIER. Antwoord A is goed.

C verwijst naar de *odds*-ratio die staat bij 'SES: laag' in de *Risk Estimate*-tabel. Dit is de *Odds Ratio for KEUKENPAPIER*. Het verband dat C suggereert is echter onzinnig: als we alleen naar mensen met een lage SES kijken, is SES toch geen

variabele meer? Verbanden kunnen alleen bestaan tussen variabelen. Als je zegt: 'arme mensen gebruiken vaker keukenpapier…' bedoel je in feite '…dan rijke mensen'. In dat geval kijk je toch ook naar rijke mensen! De *odds*-ratio die hier staat gerapporteerd, is in feite de *odds*-ratio voor het verband tussen GELUKKIG en KEUKENPAPIER, bij mensen met een lage SES. En kijk eens aan: de waarde 0,813 heb je zelf uitgerekend in opgave 1.

4. B

Opnieuw is SES de controlevariabele; het effect van deze variabele kan niet worden getoetst. In het *Total*-blok worden arme en rijke mensen op één hoop gegooid (we tellen ze niet op, maar we negeren simpelweg hun SES) en wordt de variabele dus in feite zelfs in de prullenbak gegooid. In dat geval moet het effect van KEUKENPAPIER wel getoetst worden. Aangezien we niet langer rekening houden met het feit dat de proefpersonen ook verschillen qua SES, zouden SES-effecten deze toets kunnen *confounden*.

5. B

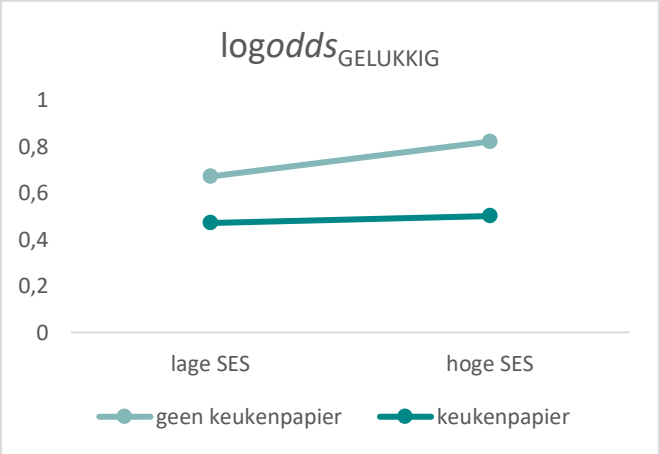
Voor de laatste keer: het effect van SES wordt in deze uitvoer niet getoetst. A valt dus af. In de steekproef zien we wel het interactie-effect dat C benoemt, maar dit effect is niet significant (*Tests of Homogeneity of the Odds Ratio*: $p = 0{,}791$). Daarmee blijft B over, en inderdaad: de *Tests of Conditional Independence* zijn niet significant ($p = 0{,}222$), dus KEUKENPAPIER heeft niet aantoonbaar een effect op GELUKKIG.

6.

Hier nog eens de kruistabel, nu met alle *odds* en log*odds*:

	lage SES		**hoge SES**	
	geen keukenpapier	**keukenpapier**	**geen keukenpapier**	**keukenpapier**
ongelukkig	49	32	40	26
gelukkig	96	51	91	43
odds	$\left(\frac{96}{49}\right) = 1{,}96$	$\left(\frac{51}{32}\right) = 1{,}59$	$\left(\frac{91}{40}\right) = 2{,}28$	$\left(\frac{43}{26}\right) = 1{,}65$
log *odds*	$\mathbf{0,67}$	$\mathbf{0,47}$	$\mathbf{0,82}$	$\mathbf{0,50}$

Van deze log*odds* kunnen we de volgende grafiek maken:



De stijging van de lijn is telkens gelijk aan de natuurlijke logaritme van de *odds*-ratio (die we in opgave 1c hebben berekend):

- Geen keukenpapier: $\ln OR = \ln 1{,}16 = 0{,}15 = 0{,}82 - 0{,}67$
- Keukenpapier: $\ln OR = \ln 1{,}04 = 0{,}03 = 0{,}50 - 0{,}47$

HOOFDSTUK 21 ENKELVOUDIGE REGRESSIE

1. B

Deze persoon zorgt ervoor dat de scores over het algemeen slechter passen bij de (vertekende) lijn, wat zorgt voor grotere voorspellingsfouten. Zulke voorspellingsfouten vormen onverklaarde variatie (waarom wijkt je woordgebruik af van de voorspelling?) en de correlatie wordt kleiner. We kunnen door deze uitschieter dus de variatie in het gebruik van Engelse woorden minder goed verklaren.

2. A

Ook de *Model ANOVA* kan worden gebruikt om te bepalen of het lezen van Nederlandse boeken een effect heeft; hij levert in dit enkelvoudige regressiemodel exact hetzelfde resultaat op als de *Coefficients*-t-toets! En hij blijkt significant ($p = 0{,}029$). Dus ja, het lezen van Nederlandse boeken lijkt een effect te hebben. De score van de proefpersoon gaat inderdaad achteruit (zie de helling b_1), maar dat was ook de bedoeling: het gebruik van Engelse termen en anglicismen daalt!

3.

Dit design is <u>observationeel</u>. Ferry Ironisch heeft zijn proefpersonen niet willekeurig de opdracht gegeven weinig of juist veel Nederlandstalige boeken te lezen. Als degenen die meer lezen minder Engelse termen en anglicismen gebruiken, kan dat dus met <u>andere</u> factoren te maken hebben. Het risico op *confounding* is reëel!

4.

Voor wie het handboek heeft gelezen: <u>foutje</u>! Het aantal vrijheidsgraden bedraagt $df = N - 2 = 25$. Mijn excuses.

De standaardfout is

$$s_b = \frac{s_{est}}{\sqrt{\sum_i (X_i - \bar{X})^2}} = \frac{15{,}296}{\sqrt{178{,}74}} = 1{,}144$$

Voor het 95%-betrouwbaarheidsinterval moeten we eerst de kritieke t-waarde opzoeken in de appendix. Bij 25 vrijheidsgraden blijkt deze gelijk aan 2,060. Het interval wordt daarmee

$$b \pm T^* s_b$$
$$-2{,}656 \pm 2{,}060 * 1{,}144$$
$$-2{,}656 \pm 2{,}357$$
$$[\,-5{,}013\,;\ -0{,}299\,]$$

Op een minuscuul decimaaltje na (een afrondingsfout) klopt dit met het interval in de SPSS-uitvoer.

HOOFDSTUK 22 MULTIPELE REGRESSIE

1.

a) $\hat{Y} = 20{,}769 + 0{,}195 * agressie - 0{,}431 * druk - 0{,}141 * AAAH!$

b) Dit is de voorspelde cortisolconcentratie (20,769 microgram per deciliter) voor een persoon die totaal niet agressief is (0), een douche zonder enige waterdruk heeft (0) en 0 graden aan de knop kan draaien voordat zijn water van bibberkoud naar kokendheet gaat. Zou zo iemand bestaan? Zo ja, dan mag hij of zij wel eens een nieuwe douche kopen.

c) Als een persoon 1 punt hoger scoort op de agressieschaal, zal zijn cortisolconcentratie naar verwachting stijgen met 0,195 microgram per deciliter. Dit is de verwachte stijging onder constanthouding van druk en AAAH!, dus als we ervan uitgaan dat deze iets agressievere persoon wel dezelfde waterdruk heeft in zijn douche en even ver moet draaien om de temperatuur drastisch te veranderen.

2. C

De ongestandaardiseerde regressiecoëfficiënten zijn schaalafhankelijk. Kijk dus in plaats hiervan naar de bèta's. Daar blijkt AAAH! de bèta te hebben die het verst afwijkt van nul.

3. B

De predictor druk heeft een tolerantiewaarde die bijna gelijk is aan 1: hij hangt dus vrijwel niet samen met agressie en AAAH!. Agressie hangt wel substantieel samen met de andere twee predictoren, want aangezien de tolerantie 0,608 is, is de proportie verklaarde variatie gelijk aan $1 - 0{,}608 = 0{,}392$. Deze samenhang moet dan wel met AAAH! zijn.

4.

a) Uitschieters in de Y-richting: ja, want minstens één *Studentized Residual* is groter dan 3.
Uitschieters in de X-richting: nee, want de maximaal toegestane *Centered Leverage Value* is $\frac{3(p+1)}{N} = \frac{3*4}{30} = 0{,}40$ en de grootste *CLV* is 0,25.
Influential cases: ja, want minstens één *Cook's Distance* is groter dan 1.

b) We hebben minstens één extreme voorspellingsfout: een persoon die een veel hoger cortisolniveau heeft dan we op basis van zijn agressie, waterdruk en AAAH! zouden voorspellen. Getuige de *Cook's Distance* trekt deze persoon de regressielijn naar zich toe; hij vertekent de analyse. (Opmerking: ik vermoed dat het om dezelfde persoon gaat.)

c) Het verband tussen druk en cortisol lijkt niet lineair… het is eerder <u>kwadratisch</u>! Dat is inhoudelijk ook veel logischer: een beetje druk op het water is prettig, maar een keiharde straal werkt juist weer stressverhogend. Hierdoor is het verband tussen druk en cortisol niet helemaal goed gemodelleerd in de regressieanalyse. Enige vertekening is het resultaat: de personen rechts bovenin trekken de regressielijn naar zich toe.

5. B

Agressie heeft totaal geen aantoonbaar effect op het cortisolniveau, dus de variatie die agressie verklaart op steekproefniveau is hoogstwaarschijnlijk toevallig. Dat betekent dat R^2 niet significant daalt als we agressie niet meenemen. Door alleen de effecten van druk en AAAH! te bekijken, beschrijven we de populatie (de werkelijkheid) even goed.

6. B

Met een *backward*-procedure kom je uit bij model 2. Agressie blijkt in model 3 namelijk niet significant ($p > 0{,}10$) en kan weg. In model 2 is druk nog steeds niet helemaal significant, maar om Type II-fouten te voorkomen, halen we in *backward* pas predictoren uit het model als de p-waarde groter is dan 0,10. Daarom wordt het model met druk en AAAH! het eindmodel. De conclusie die daaruit volgt, staat in antwoord B beschreven.

HOOFDSTUK 23 LOGISTISCHE REGRESSIE

Ter verduidelijking van de uitwerking volgt hier de dataset; die zou je ook in je eigen onderzoek natuurlijk tot je beschikking hebben.

	merel		blinde vink	
	normaal raam	geblindeerd raam	normaal raam	geblindeerd raam
botsing: nee	38	48	19	23
botsing: ja	22	12	27	23
	60	60	46	46

1.

a) Ja: de toetsen *Step* en *Block Chi-square* zijn significant ($p = 0{,}049$) in de *Omnibus*-tabel, wat aangeeft dat model 1 een beter beeld van de populatie geeft dan het predictorloze model 0. Ook is de regressiecoëfficiënt (β) van RAAM significant ($p = 0{,}050$) in de *Variables in the Equation*-tabel, waarmee een effect van het raam wordt aangetoond.

b) Mogelijk wordt het resultaat van model 1 vertekend door *confounding* en/of interactie. Het effect van het raam zou bijvoorbeeld kunnen verschillen afhankelijk van de VOGEL (de tweede predictor), en dat is interactie. Daarnaast is de p-waarde van het RAAM-effect gelijk aan het significantieniveau α, en dus is de toets slechts héél krapjes significant: het risico bestaat dat we een Type I-fout maken (dat wil zeggen dat we de nulhypothese van geen effect onterecht verwerpen).

c) 0 betekent een normaal raam, 1 een geblindeerd raam. De β van RAAM geeft aan hoe sterk de log*odds* op een botsing zullen stijgen als we op de predictor 1 punt naar boven gaan, dus van een normaal raam naar een geblindeerd raam. De log*odds* dalen in dat geval met 0,556. Bij een geblindeerd raam zijn de log*odds* op een botsing dus lager, en is de kans op een botsing derhalve kleiner. Dat maakt het normale raam voor vogels het gevaarlijkst.

2. A

Het feit dat de interactieterm in model 3 niet nul is, betekent nog niet dat zich ook in de populatie een interactie-effect voordoet; om dat te bewijzen, hebben we de statistische toets nodig. B valt dus af.
In feite is het waar dat in model 3 een simpel effect van RAAM wordt getoetst (bij merels) en een simpel effect van VOGEL (bij een normaal raam). Zie de opmerking hierover in paragraaf 23.5 (*Dichotome predictoren plus interactie: steekproefresultaten*). Deze simpele effecten worden ook aangetoond (de toetsen zijn significant). De aanwezigheid van simpele effecten zegt echter niets over interactie. Bij non-interactie zijn alle simpele effecten van een predictor eenvoudigweg hetzelfde, en dus heeft het meer zin om naar het hoofdeffect te kijken. Ook C is verkeerd.
Om toch vast te stellen of er een interactie-effect is, kunnen we ook gebruikmaken van de *Hosmer and Lemeshow Test* bij model 2. De nulhypothese dat dit model compleet is – en dat we dus geen interactieterm hoeven toe te voegen – kan helemaal niet worden verworpen ($p = 0{,}713$). Daarom zal professor McDuck in blok 3 waarschijnlijk geen significant interactie-effect hebben gevonden.

3. C

$$\log odds = b_0 + b_1 * RAAM + b_2 * VOGEL + b_3 * RAAM * VOGEL$$

Als we de b's invullen, krijgen we:

$$\log odds = -0{,}547 - 0{,}840 * RAAM + 0{,}898 * VOGEL + 0{,}488 * RAAM * VOGEL$$

Nu kunnen we voor elke situatie de log*odds* op een botsing uitrekenen. Hoe hoger de log*odds*, des te groter de kans op een botsing.

SITUATIE	RAAM	VOGEL	$\log odds =$	$\log odds =$
Normaal raam, merel	0	0	b_0	$-0,547$
Geblindeerd raam, merel	1	0	$b_0 + b_1$	$-1,386$
Normaal raam, blinde vink	0	1	$b_0 + b_2$	$0,351$
Geblindeerd raam, blinde vink	1	1	$b_0 + b_1 + b_2 + b_3$	0

Van de drie antwoordmogelijkheden heeft de blinde vink bij een geblindeerd raam de grootste log*odds*. De blinde vink bij een normaal raam staat eigenlijk bovenaan, maar die kunnen we nu niet kiezen.

4. B

De natuurlijke logaritme van een *odds*-ratio is gelijk aan een b of een combinatie van b's, dus we moeten op zoek naar de juiste e-macht van een (combinatie van) b's om de gevraagde *odds*-ratio te vinden. Je kunt grafiekjes maken van de log*odds* – dat is het meest overzichtelijk – maar het kan ook met de tabel hierboven:

- De onderste twee rijen gaan over de blinde vinken;
- Het verschil tussen die twee rijen is een normaal versus een geblindeerd raam;

- Het verschil in b's is $b_1 + b_3$. Dit is dus het (simpele) effect van het raam bij blinde vinken;
- De e-macht van deze b's is de *odds*-ratio $e^{b_1+b_3} = e^{-0,840+0,488} = \mathbf{0,703}$.

Pak niet de e-macht van enkel b_1; dit is het (simpele) effect van het raam bij merels! Je zult dan bij antwoord A uitkomen (dat ook in de uitvoer staat als *Exp(B)*) en dat is onjuist.

5. A

Model 3 is niet spaarzaam genoeg, zo stelden we al eerder vast; er is geen interactie-effect, dus C (de definitie van een interactie-effect) kunnen we schrappen. A en B beschrijven allebei een simpel effect, maar omdat er geen interactie is, is het effect van het raam voor alle vogels hetzelfde. Bekijk model 2: de b van RAAM is -0,599, dus negatief. Het geblindeerde raam levert dus de laagste log*odds* op een botsing op. Zowel merels als blinde vinken botsen minder vaak tegen een geblindeerd raam dan tegen een normaal raam.

6. B

De *Mantel-Haenszel Common Odds Ratio Estimate* is een soort gewogen gemiddelde van de simpele effecten. Door een gemiddelde te berekenen, ga je ervan uit dat de simpele effecten puur door toeval verschillen en dat er geen interactie is. We zoeken dus naar het gecorrigeerde hoofdeffect van VOGEL, in een model zonder interactieterm. Dit is natuurlijk model 2. De *odds*-ratio voor het verband tussen BOTSING en VOGEL vinden we onder *Exp(B)*: 3,083.

7.

a) $OR = e^{0,116} = 1,12$
b) Ze worden 1,12 keer zo groot. Ik vraag gewoon naar de inhoudelijke betekenis van de *odds*-ratio! ☺
c) Dit is een stijging van de log*odds* met $4 * b$, dus $OR = e^{4*0,116} = 1,59$.

HOOFDSTUK 24 BETROUWBAARHEID

1.

a) Ik vind van niet: ik geloof dat alle items meten hoe de wetenschapper de balans benadert tussen toegankelijk onderwijs enerzijds, en serieus en inhoudelijk correct onderwijs anderzijds. Behalve item IV: dat meet eerder hoe leuk de wetenschappers hun vak vinden.
b) Ja: kijk maar naar de *Item-Total Statistics*. De *Corrected Item-Total Correlation* van item IV is heel laag. En als we item IV verwijderden, zou Cronbachs alfa stijgen *(Cronbach's Alpha if Item Deleted)*.
c) De items zijn niet unidimensionaal (ze meten gezamenlijk niet één persoonskenmerk) en dus ook niet parallel.
d) Als gevolg hiervan zullen de items onderling minder sterk correleren. We gebruiken die correlaties als schatters van de betrouwbaarheid van de items, dus die betrouwbaarheid zullen we onderschatten. We mogen erop rekenen dat de items betrouwbaarder zijn dan we in de analyse terugzien.

2. C

Bekijk de *Item Statistics*. Aangezien het gemiddelde op item IV beduidend hoger is, meet dit item waarschijnlijk niet dezelfde ware score T. De parallelliteitsassumptie is in dat geval geschonden; vandaar argument A.
Ook de standaarddeviatie op item IV is erg gering (argument B). Bijna alle wetenschappers hebben dus aangegeven dat ze het eens zijn met de stelling: natuurlijk zijn moleculen onwijs cool. Had je een ander antwoord mogen verwachten van een chemicus? Dit item slaagt er dan ook niet echt in om de deelnemers van elkaar te onderscheiden, om de meer subtiele nuances op te pikken waarin ze verschillen qua passie voor hun vak. Dat maakt het item tot een vrij nutteloos psychometrisch instrument.

3. B

Cronbachs alfa *(Reliability Statistics)* is te laag; hij zou eerder 0,80 moeten zijn. Dit zouden de onderzoekers kunnen bewerkstelligen door de vragenlijst uit te breiden met meer (parallelle) items. De huidige items – nummer IV uitgezonderd – lijken redelijk parallel, dus gezamenlijk kunnen ze een enigszins betrouwbaar beeld geven van hetzelfde persoonskenmerk: de didactische overtuigingen van een wetenschapper. Helemaal onbruikbaar is de vragenlijst dus niet.

4. C

Check de *Item-Total Statistics: Cronbach's Alpha if Item Deleted*. Zonder item IV wordt Cronbachs alfa gelijk aan 0,629.
Nu zouden we eerst de geschatte betrouwbaarheid per item kunnen uitrekenen door de Spearman-Brown-formule voor Cronbach's alfa op te lossen, en deze formule vervolgens opnieuw in te vullen met $k = 10$. Dit doe ik meteen hieronder, maar dat doe ik alleen om te laten zien dat er een snellere methode is – een die ook klopt!
Met $k = 4$:

$$\alpha = \frac{k * \bar{r}_{item-item'}}{1 + (k-1)\bar{r}_{item-item'}}$$
$$0{,}629 = \frac{4 * \bar{r}_{item-item'}}{1 + 3 * \bar{r}_{item-item'}}$$
$$0{,}629 * (1 + 3 * \bar{r}_{item-item'}) = 4 * \bar{r}_{item-item'}$$
$$0{,}629 + 1{,}887 * \bar{r}_{item-item'} = 4 * \bar{r}_{item-item'}$$
$$0{,}629 = 2{,}113 * \bar{r}_{item-item'}$$
$$\bar{r}_{item-item'} = \frac{0{,}629}{2{,}113} = 0{,}298$$

Met $k = 10$:

$$\alpha = \frac{k * \bar{r}_{item-item'}}{1 + (k-1)\bar{r}_{item-item'}} = \frac{10 * 0{,}298}{1 + 9 * 0{,}298} = \mathbf{0{,}809}$$

De snelste en gemakkelijkste methode is echter om te stellen dat het aantal items (k) zou stijgen van 4 naar 10. Dit staat gelijk aan een stijging van 1 naar 2,5 $\left(\frac{10}{4}\right)$. Bizar genoeg kun je nu gewoon in de Spearman-Brown-formule $k = 2{,}5$ en de huidige betrouwbaarheid 0,629 invullen!

$$\boldsymbol{\rho_{10\ items}} = \frac{k * \rho_{4\ items}}{1 + (k-1)\rho_{4\ items}} = \frac{2{,}5 * 0{,}629}{1 + 1{,}5 * 0{,}629} = \mathbf{0{,}809}$$

5. C

De gemiddelde somscore van de twintig deelnemers is 16,45 (*Scale Statistics: Mean*).
A is sowieso het verkeerde antwoord: Bacarra's hogere (gemeten) score zou ook door meetfouten kunnen komen. We moeten dus een **betrouwbaarheidsinterval** opstellen voor het ware verschil tussen de twee somscores.
De somscores van de deelnemers hebben een variantie van 10,576 (zie de *Scale Statistics: Variance*). Deze gemeten variantie komt voor een groot deel door meetfouten: Cronbachs alfa is 0,569, dus 56,9% van de gemeten spreiding bestaat uit ware verschillen tussen de deelnemers, en $100\% - 56{,}9\% = 43{,}1\%$ bestaat uit verschillen door meetfouten.
De meetfoutvariantie van een losse somscore is daarom (naar schatting)

$$\sigma^2_{e_{som}} = (1 - \rho_{SS'}) * \sigma^2_{som} = (1 - 0{,}569) * 10{,}576 = 0{,}431 * 10{,}576 = 4{,}558$$

Echter, de meetfoutvariantie van het <u>verschil</u> tussen twee somscores is nog eens tweemaal zo groot:

$$\sigma^2_{e_{somverschil}} = 2 * 4{,}558 = 9{,}116$$

De geschatte standaardmeetfout is gelijk aan de wortel hiervan:

$$SEM = \sqrt{\sigma^2_{e_{somverschil}}} = \sqrt{9{,}116} = 3{,}02$$

Aannemend dat de meetfouten normaal verdeeld zijn en dat hun gemiddelde 0 is, mogen we stellen dat 95% van alle meetfouten die er op het verschil tussen twee somscores gemaakt worden, zich bevindt tussen

$$[\,-2 * 3{,}02\,;\ 2 * 3{,}02\,] = [\,-6{,}04\,;\ 6{,}04\,]$$

Dat betekent dat de meetfout op het verschil tussen Bacarra's en Spencers somscore waarschijnlijk (met 95% zekerheid) niet groter is dan 6,04. Het ware verschil kan dus overal liggen tussen

$$23 - 19 \pm 6{,}04$$

$$4 \pm 6{,}04$$

$$[\mathbf{-2{,}04; 10{,}04}]$$

Het ware verschil kan blijkbaar ook 0 punten bedragen. Hiroshima en zijn collega's kunnen dus niet aantonen dat Bacarra progressiever is dan Spencer. Wellicht was dit wel gelukt met een betrouwbaardere vragenlijst.

6.

a) Hiervoor gebruiken we de attenuatieformule. De betrouwbaarheid van de vragenlijst is nog steeds naar schatting 0,569). Beschouw de somscores op de vragenlijst als X en het gemiddelde oordeel over het handboek als Y.

$$\boldsymbol{\rho_{T_X T_Y}} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}} = \frac{0{,}32}{\sqrt{0{,}569 * 0{,}88}} = \mathbf{0{,}45}$$

b) Dat niet per se: de correlatie is tenslotte niet 1. Wel doet zich een positieve trend voor: handboeken die geschreven zijn door een progressieve hoogleraar, worden door studenten over het algemeen wat beter bevonden.

HOOFDSTUK 25 OVEREENSTEMMING

1.
Om te beginnen,

$$A_o = 1 + 2 + 5 + 3 + 3 = 14$$

Als we slim zijn, beperken we ons tot het berekenen van de verwachte overeenstemmingen (de diagonaal). Dit zijn de betreffende *Expected Counts*:

- $EC_{rood,rood} = \frac{5*10}{40} = 1{,}25$
- $EC_{oranje,oranje} = \frac{10*5}{40} = 1{,}25$
- $EC_{groen,groen} = \frac{5*15}{40} = 1{,}875$
- $EC_{blauw,blauw} = \frac{15*5}{40} = 1{,}875$
- $EC_{paars,paars} = \frac{5*5}{40} = 0{,}625$

In de kruistabel:

Expected Counts		ARYOO					
		rood	**oranje**	**groen**	**blauw**	**paars**	
NUTS	**rood**	1,25					**5**
	oranje		1,25				**10**
	groen			1,875			**5**
	blauw				1,875		**15**
	paars					0,625	**5**
		10	**5**	**15**	**5**	**5**	**40**

De verwachte overeenstemming is dus

$$A_e = 1{,}25 + 1{,}25 + 1{,}875 + 1{,}875 + 0{,}625 = 6{,}875$$

Kortom, kappa wordt

$$\kappa = \frac{A_o - A_e}{N - A_e} = \frac{14 - 6{,}875}{40 - 6{,}875} = 0{,}22$$

2. C
Immers, kappa is ronduit bedroevend. ☺

3. C
De themakleur is een <u>nominale</u> variabele: we mogen de kleuren zo in een andere volgorde zetten, en dan zou een gewogen kappa telkens veranderen afhankelijk van de volgorde.

4.
a) Volg de formules in de theorie en je komt uit op de volgende wegingscoëfficiënten:

		ARYOO (*i*)				
		Kleingeld (1)	**Bescheiden (2)**	**Forse som (3)**	**Fortuin (4)**	
NUTS (*j*)	**Kleingeld (1)**	1	0,67	0,33	0	
	Bescheiden (2)	0,67	1	0,67	0,33	
	Forse som (3)	0,33	0,67	1	0,67	
	Fortuin (4)	0	0,33	0,67	1	

Als het goed is, zie je de overeenstemming dus <u>in gelijkmatige stappen</u> (dat is de definitie van lineair!) afnemen van 1 tot 0.
Was dat al het gruwelijke rekenwerk? Grotendeels:

$$A_o = (10 + 4 + 5 + 9) * 1 + (0 + 0 + 5 + 5 + 0 + 0) * 0{,}67 + (0 + 1 + 0 + 1) * 0{,}33 + (0 + 0) * 0 = 28 * 1 + 10 * 0{,}67 + 2 * 0{,}33 + 0 * 0 = 35{,}33$$

$$A_e = (3{,}75 + 1{,}25 + 1{,}25 + 3{,}75) * 1 + (1{,}25 + 1{,}25 + 3{,}75 + 3{,}75 + 1{,}25 + 1{,}25) * 0{,}67 + (1{,}25 + 3{,}75 + 3{,}75 + 1{,}25) * 0{,}33 + (3{,}75 + 3{,}75) * 0 = 10 * 1 + 12{,}50 * 0{,}67 + 10 * 0{,}33 + 7{,}50 * 0 = 21{,}67$$

$$\kappa = \frac{35{,}33 - 21{,}67}{40 - 21{,}67} = 0{,}75$$

Niet slecht!

b) Pff… ☹ eh, ik bedoel – NATUURLIJK kunnen we dat! Helemaal te gek, joh!

		ARYOO (*i*)				
		Kleingeld (1)	**Bescheiden (2)**	**Forse som (3)**	**Fortuin (4)**	
NUTS (*j*)	**Kleingeld (1)**	1	0,89	0,56	0	
	Bescheiden (2)	0,89	1	0,89	0,56	
	Forse som (3)	0,56	0,89	1	0,89	
	Fortuin (4)	0	0,56	0,89	1	

HOOFDSTUK 26 MODERNE PSYCHOMETRIE

1. B

De assumptie van unidimensionaliteit (A) geldt voor de klassieke én de moderne modellen. Er zijn ook moderne modellen voor polytome en kwantitatieve items in ontwikkeling, dus ook C is fout (maar in *Piraten, perziken en p-waarden* bespreken we alleen dichotome items in de context van de moderne itemresponstheorie). Antwoord B gaat over de assumptie van parallelle items: het klassieke model neemt dit aan, de moderne modellen niet.

2. C

Verschillende a's komen voort uit verschillende betrouwbaarheden; de meetfoutvariantie verschilt dan per item. Dit schendt de assumptie van parallelliteit. Items die niet parallel zijn, zullen minder met elkaar correleren. Aangezien we die correlatie gebruiken om de betrouwbaarheid van de items te meten, zullen we de betrouwbaarheid van de items onderschatten – en daarmee ook de betrouwbaarheid van de complete vragenlijst.
Ook verschillende b's betekenen een schending van parallelliteit: items met verschillende moeilijkheidsgraden zijn niet parallel. Dat betekent dat beide uitspraken correct zijn.

3. A

Het item met de meest rechtse *ICC* heeft de hoogste moeilijkheidsgraad: de hoogste b. De moeilijkheidsgraad van een item kunnen we schatten met de item-p-waarde: de proportie enen (de proportie personen die het goed hadden, die 'ja' invulden of 'eens'). Deze vinden we in de *Item Statistics: Mean*. Hoe hoger die proportie, des te gemakkelijker het item en dus te lager de b. We zoeken echter naar het moeilijkste item, dus naar het exemplaar met de kleinste item-p-waarde. Dit is item I.

4. A

Het item met de steilste *ICC* heeft de hoogste a. Aangezien de a van een item rechtstreeks geschat wordt uit zijn item-restcorrelatie, moeten we bekijken welk item de hoogste *Corrected Item-Total Correlation* heeft (zie *Item-Total Statistics*). Dit blijkt item VI te zijn.

5. A

Ten eerste: waarom niet B? Zeiden we in hoofstuk 24 niet een keer dat proefpersonen divers moeten scoren op een item? Immers, als iedereen hetzelfde scoort, kan het item niemand onderscheiden. Het probleem is dat de spreiding op item VIII (uitgedrukt door de grootste standaardafwijking) ook grotendeels door meetfouten kan komen. In dat geval zou een item met minder spreiding, grotendeels door ware verschillen, nuttiger zijn.
Het item met het hoogste onderscheidingsvermogen is volgens klassieke modellen het item met de hoogste betrouwbaarheid. Een betrouwbaar item, dat steeds ongeveer dezelfde uitslag geeft bij dezelfde proefpersoon, zou een goede kans moeten maken om met succes te bepalen welke proefpersoon hoog scoort en welke laag. We zoeken dus naar het item met de hoogste item-restcorrelatie, en dat is, zoals we bij opgave 4 al vaststelden, item VI.
Ook de moeilijkheidsgraad (b) heeft weliswaar iets met de informativiteit van een item te maken, maar hierover kan geen op zichzelf staande uitspraak worden gedaan. Zie opgave 6.

6.

a) Opmerking voor gebruikers van het handboek, editie 2015: mijn excuses, de tabel in het boek klopte niet! Gelieve de onderstaande lijst met a's en b's te gebruiken en de opgave nog een keer te maken. Beschouw het maar als een extra oefening ☺

Item	θ	a	b	$a(\theta - b)$	P	$I(\theta)$
I	0	1,5	2	-3	0,05	0,1069
II	0	0,5	0	0	0,50	0,0625
III	0	0,75	2	-1,5	0,18	0,0830
V	0	0,75	2	-1,5	0,18	0,0830
VI	0	1,5	0	0	0,50	0,5625
VII	0	0,75	2	-1,5	0,18	0,0830
VIII	0	1,5	-1	1,5	0,82	0,3321

b) Ja, het antwoord wordt bevestigd! Maar dat had niet zo hoeven te zijn als Hiroshima bijvoorbeeld extreem progressieve docenten had willen scheiden van de rest. In dat geval had hij de wetenschappers moeilijkere items moeten voorleggen. (De klassieke testleer, toegepast in Bijlage 26, houdt geen rekening met het effect van verschillende θ's.)

c) Tel alle iteminformaties op: $I_{test}(\theta = 0) = 1{,}313$.

d) De testinformatie van zo'n test zou gelijk zijn aan $I_{test}(\theta = 0) = 3 * 0{,}0625 + 3 * 0{,}5625 = 1{,}875$. Dat maakt een homogene test in dit geval informatiever dan Hiroshima's huidige exemplaar.

HOOFDSTUK 27 FACTORANALYSE

1. B

Kaiser-criterium: 3 factoren
Screeplot: 3 factoren (C is dus fout)
Lage residuele correlaties: staat niet expliciet in de uitvoer
Maximum Likelihood: incompleet. Nodig is ook de *Goodness of Fit*-toets bij 2 factoren en misschien zelfs bij 1 factor, om te kijken welk model het spaarzaamst is én niet wordt verworpen. We kunnen dan ook niet met zekerheid zeggen dat alle criteria het met elkaar eens zijn.

2. A

Gebruik niet de *Initial Eigenvalue* in de *Total Variance Explained*-tabel! Hier staan de eigenwaarden van PCA. Die van PFA *(Extraction Sums of Squared Loadings)* zijn weggelaten. In dat geval moeten we de eigenwaarde van Factor 3 maar berekenen door alle ladingen te kwadrateren en op te tellen. Je vindt ze in de *Factor Matrix* (niet de geroteerde oplossing gebruiken!):

$$eigenwaarde_{F3} = \sum \lambda_j^2 =$$
$$0{,}484^2 + 0{,}095^2 + 0{,}593^2 + (-0{,}033)^2 + (-0{,}278)^2 +$$
$$0{,}052^2 + (-0{,}321)^2 + (-0{,}285)^2 + 0{,}209^2 = \mathbf{0{,}904}$$

Het K1-criterium moet worden toegepast middels de eigenwaarden van PCA. Bij PCA zou Factor 3 een voldoende hoge eigenwaarde hebben gehad, namelijk 1,279.

3. A

De communaliteit beschrijft hoe goed de variantie in de losse items wordt verklaard; het gaat ons echter om een verklaring van de covariantie ofwel de correlatie tussen item 4 en 6. Daarom moeten we nagaan hoe goed die correlatie van 0,699 door de factoren wordt gereproduceerd. Gebruik de *Factor Matrix* om alle ladingen op te zoeken.

$$r_{reprod} = \lambda_{4,1}\lambda_{6,1} + \lambda_{4,2}\lambda_{6,2} + \lambda_{4,3}\lambda_{6,3} =$$
$$0{,}674 * 0{,}652 + (-0{,}552) * (-0{,}426) + (-0{,}033) * 0{,}052 = 0{,}673$$

De residuele correlatie is het verschil tussen de echte correlatie tussen item 4 en 6, en de gereproduceerde correlatie:

$$\boldsymbol{r_{res}} = r - r_{reprod} = 0{,}699 - 0{,}673 = \mathbf{0{,}026}$$

Dit is wat A zegt en is inderdaad een klein getal! Het verband tussen stress en sarcasme acteren wordt door een model met drie factoren erg goed verklaard.

4. C

Het gaat hier om de communaliteit van item 5. Deze kunnen we opzoeken: hij is 0,906. Ruim 90% van de spreiding in de individuele scores op item 5 wordt dus verklaard door de factoren. Daarmee blijft er $1 - 0{,}906 = 0{,}094$ aan onverklaarde spreiding over: de uniciteit, een kleine 10%. Dit is vrij weinig! ☺

5. A

Voordat er geroteerd werd, laadden veel items nog hoog op diverse factoren *(Factor Matrix)*. Dat maakte het erg moeilijk om te bepalen welke factor elk item nu precies meet. De simpele structuur die na scheve rotatie ontstaat *(Pattern Matrix)*, is een stuk duidelijker: er is nu telkens één factor waarop elk item veel hoger laadt (A is goed). Het kan dan best zijn dat een item toch bij een andere factor blijkt te horen dan voor de rotatie leek (C is fout). Meestal is een scheve rotatie beter dan een orthogonale rotatie; in elk geval geldt niet andersom (B is fout).

6.

a) Check de *Pattern Matrix*:
   - Item 1 (blijdschap) en 3 (enthousiasme) laden het hoogst op Factor 3. Deze factor lijkt zoiets als uitbundig acteerwerk; het krachtig uitdragen van positieve emoties. Dit is slechts mijn eigen omschrijving. Verzin gerust een betere!
   - Item 2 (aarzeling), 4 (stress), 6 (sarcasme) en 9 (medeleven) laden het hoogst op Factor 1. De overeenkomst tussen deze items lijkt mij dat ze allemaal meer subtiel acteerwerk vereisen: er komt geen groot drama bij kijken, eerder nuance.
   - Item 5 (wanhoop), 7 (woede) en 8 (verdriet) ten slotte laden het hoogst op Factor 2. Hier gaat het duidelijk om dramatisch acteerwerk: het vertolken van krachtige negatieve emoties.

b) De *Factor Correlation Matrix* geeft aan dat sommige factoren wel wat met elkaar te maken hebben, maar niet heel veel. Zo is de correlatie tussen Factor 1 en 2 gelijk aan 0,285: wie subtiel kan acteren, kan soms ook dramatisch acteren (maar zeker niet altijd). De correlatie tussen Factor 2 (subtiel acteren) en 3 (uitbundig acteren) is vrijwel hetzelfde: 0,312. De correlatie tussen Factor 1 en 3 daarentegen blijkt bijna gelijk aan 0 (namelijk 0,052). Van een goede uitbundige acteur valt dus helemaal niet te voorspellen of deze ook een goede dramaqueen zal zijn, en evenmin andersom. Het zou kunnen, maar het hoeft niet. Deze talenten hangen waarschijnlijk niet met elkaar samen.

c) Het nadeel: met minder items wordt een meting van acteertalent minder betrouwbaar! Ik zou dit dus niet zonder meer aanraden.

DISCLAIMER
Wijzigingen en fouten voorbehouden.
Denk je dat je een fout hebt gevonden,
neem dan contact op met vincependers@live.nl
of met facebook.com/pppwaarden.
We verbeteren hem graag! ☺