Hoofdstuk 1

1. Welke van de onderstaande maten kan worden berekend uit de *five-number summary* (vijf-getallen-samenvatting)?

A. Het gemiddelde
B. De interkwartiele range
C. De standaarddeviatie
D. De variantie

2. Persoon *X* heeft veel oefententamens van statistiek gemaakt. Hierdoor begrijpt *X* de stof goed en haalt het tentamen. De variabele 'aantal uren studeren' is een voorbeeld van een

A. Afhankelijke variabele
B. Normaal verdeelde variabele
C. Onafhankelijke variabele
D. Kwalitatieve variabele

3. Een docent heeft een *stemplot* (stam-en-bladdiagram) gemaakt van het aantal punten dat iedere leerling op het tentamen statistiek (schaal 0-100) heeft gehaald. Uit het stemplot blijkt dat de modus gelijk is aan 61. Welke van de onderstaande stemplots zou hierop van toepassing kunnen zijn?

A.

3	8				
4	2	8			
5	4	5	6	7	
6	1	1	1	6	
7	3	3	8	8	
8	0	2	2	5	9
9	3	5	9		

B.

3	8				
4	2	3	8		
5	4	5	5	5	
6	0	0	1	6	
7	3	3	8	8	9
8	0	2	5		
9	3	5	9		

C. Geen van de bovenstaande stemplots zou van toepassing kunnen zijn.
D. Beide stemplots zouden van toepassing kunnen zijn.

4. Met behulp van welke figuur kun je het beste zien of de scores op een variabele normaal verdeeld zijn?

A. Q-Q plot
B. Staafdiagram
C. Tijddiagram
D. Histogram

5. Van een groep eerstejaars Psychologiestudenten zijn de tentamencijfers voor Statistische modellen 1 bekend. De *five-number summary* van deze tentamencijfers is als volgt:

4 5 6 7 9

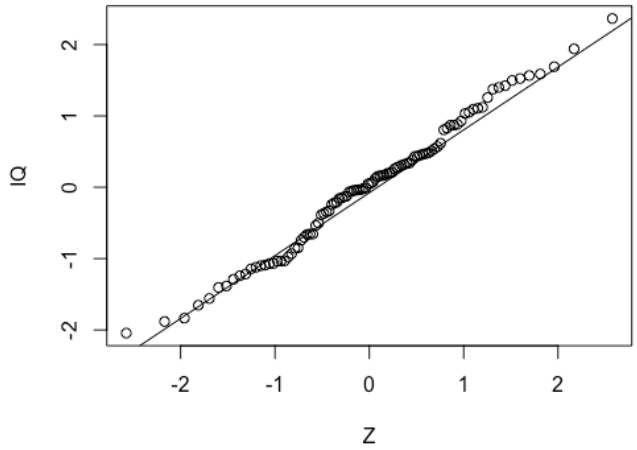
Welke bewering is waar?

A. De scores boven de modus zijn minder verspreid dan de scores onder de modus.
B. De scores boven de modus zijn meer verspreid dan de scores onder de modus.
C. De scores boven de mediaan zijn minder verspreid dan de scores onder de mediaan.
D. De scores boven de mediaan zijn meer verspreid dan de scores onder de mediaan.

6. Wat valt niet uit een boxplot af te leiden, wanneer de variabele scheef verdeeld is?

A. Het gemiddelde
B. De mediaan
C. De interkwartiele range
D. Het minimum

7. Wat voor plot staat hieronder afgebeeld?



A. Density plot
B. Normal Quantile plot
C. Line plot
D. Residual plot

8. De scores van 400 proefpersonen op een intelligentietest hebben een gemiddelde van 300 en een standaarddeviatie van 30. De onderzoeker wil de scores lineair transformeren zodat het gemiddelde 100 is en de standaarddeviatie 15. Wat moet de onderzoeker doen?

A. Alle scores delen door 2.
B. Alle scores delen door 3.
C. Alle scores delen door 2 en er 50 vanaf trekken.
D. Alle scores delen door 2 en er 100 vanaf trekken.

9. Welke van de onderstaande beweringen is/zijn waar?

I. De standaarddeviatie is resistent

II. De standaarddeviatie is nul wanneer er geen uitbijters zijn

A. Alleen bewering I is waar
B. Alleen bewering II is waar
C. Beide beweringen zijn waar
D. Beide beweringen zijn niet waar

10. De verdeling van huizenprijzen blijkt rechtsscheef verdeeld te zijn. De gemiddelde huizenprijs is 223500 euro. Dan is de mediaan

1. Lager dan 223500
2. Gelijk aan 223500
3. Hoger dan 223500
4. Daar kan op basis van deze gegevens geen uitspraak over worden gedaan

11. Van een test is bekend dat deze een gemiddelde heeft van 100 en een standaarddeviatie van 30. Een onderzoeker wil de scores zodanig transformeren, dat de standaarddeviatie 15 wordt, maar het gemiddelde gelijk blijft aan 100. Met welke van de onderstaande transformaties zal hij dit bereiken?

A. $Y = 0.50X$
B. $Y = 0.50X + 50$
C. $Y = 2X$
D. Dat is niet mogelijk

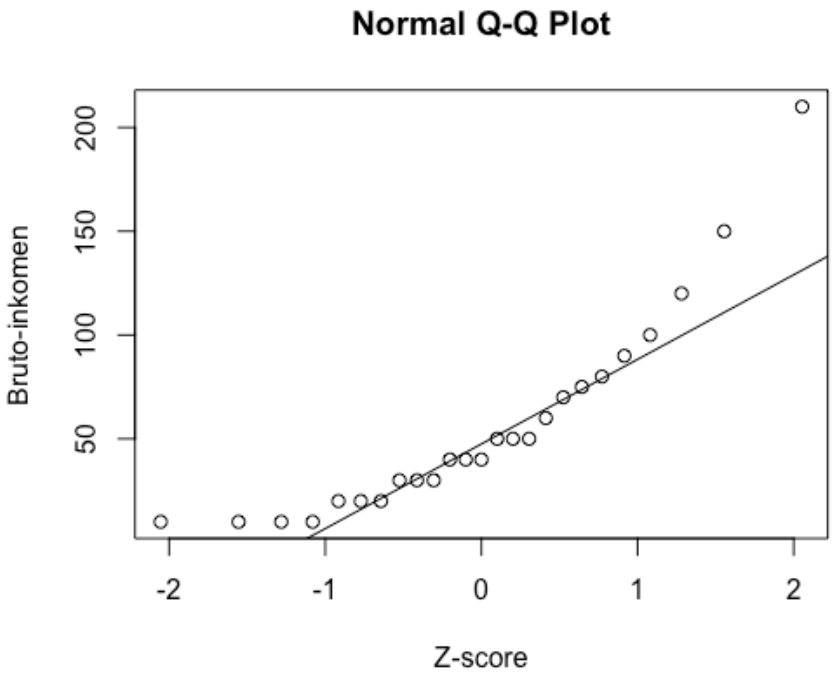
12. Gegeven is een five-number summary met de volgende getallen: 20, 25, 28, 35 en 55. Welke van de volgende scores zou volgens het 1.5*IQR-criterium als uitbijter worden aangemerkt?

A. 15
B. 55
C. Zowel 1 als 55
D. Geen van bovenstaande

13. Een onderzoeker wil zijn data set beschrijven met twee samenvattingsmaten: een centrummaat en een spreidingsmaat. Waar kan de onderzoeker het beste voor kiezen, als hij de data set wil beschrijven met zo robuust mogelijke maten?

A. Gemiddelde en standaarddeviatie
B. Gemiddelde en IQR
C. Mediaan en standaarddeviatie
D. Mediaan en IQR

14. Een onderzoeker heeft van 500 mensen gegevens verzameld over het maandelijks bruto-inkomen en de benzinekosten per maand. Op basis van de verzamelde gegevens maakt de onderzoeker het onderstaande Q-Q plot. Welke van de volgende conclusies is juist?



A. Het bruto-inkomen correleert sterk met de maandelijkse benzinekosten
B. Het bruto-inkomen lijkt normaal verdeeld
C. Het bruto-inkomen correleert niet sterk met de maandelijkse benzinekosten
D. Het bruto-inkomen lijkt niet perfect normaal verdeeld

15. Een onderzoeker heeft gegevens verzameld over de leefsituatie van studenten en deze opgedeeld in de volgende categorieën: zelfstandig (studio), samenwonend met partner, samenwonend met andere studenten (studentenhuis), bij ouders. De onderzoeker wil de verzamelde gegevens grafisch weergeven. Welke figuur kan hij hiervoor het beste gebruiken?

A. Boxplot
B. Stemplot
C. Staafdiagram
D. Spreidingsdiagram

16. In een internationaal onderzoek over meerdere landen bij mannen en vrouwen wordt gekeken in hoeverre het bruto-inkomen voorspeld kan worden aan de hand van het opleidingsniveau. Wat is hier de onafhankelijke variabele?

A. Nationaliteit
B. Geslacht
C. Bruto-inkomen
D. Opleidingsniveau

17. Wat betekent een interkwartielrange (IQR) van 16?

A. Dat de middelste 50% van de scores verspreid liggen over een schaalbreedte van 4 punten.

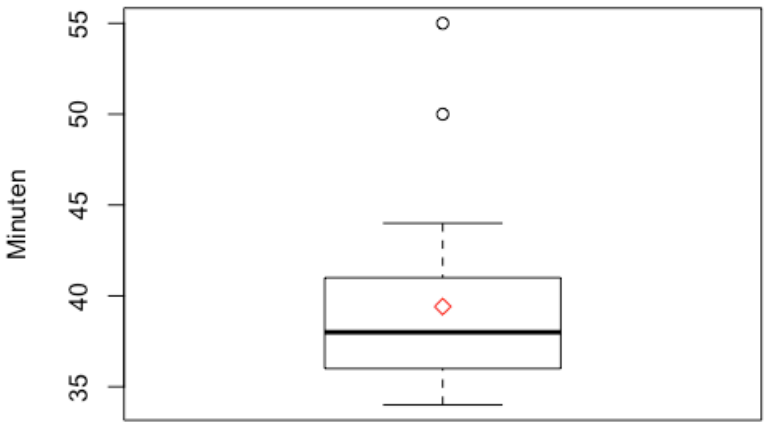
B. Dat de middelste 50% van de scores verspreid liggen over een schaalbreedte van 8 punten.
C. Dat de middelste 50% van de scores verspreid liggen over een schaalbreedte van 16 punten.
D. Dat de middelste 50% van de scores verspreid liggen over een schaalbreedte van 32 punten.

18. Van 1500 kinderen is gemeten hoe lang zij erover doen om een bepaalde tekst te schrijven. Er wordt aangenomen dat de variabele 'tijd' normaal verdeeld is in de populatie. Uit een aselecte steekproef van 2500 personen blijken 95% van de scores tussen de 5 en 9 minuten te liggen. Welke van de onderstaande uitspraken is waar?

I. De standaarddeviatie in de steekproef zal hoogstwaarschijnlijk ongeveer 1 zijn.
II. Het steekproefgemiddelde zal hoogstwaarschijnlijk ongeveer 7 zijn.

A. Alleen bewering I is waar
B. Alleen bewering II is waar
C. Beide beweringen zijn waar
D. Beide beweringen zijn niet waar

19. Afgelopen weekend vielen de bladeren weer van de boom. Dit leverde veel problemen op bij de NS. Van een aselecte steekproef van 100 personen is bekend hoeveel minuten vertraging zij dit weekend hadden bij de NS. Op basis van deze gegevens is het onderstaande boxplot opgesteld. Wat zou het vierkantje redelijkerwijs kunnen aangeven?



A. De mediaan
B. De positie van de mediaan na weglating van de uitbijters
C. De IQR
D. Het gemiddelde

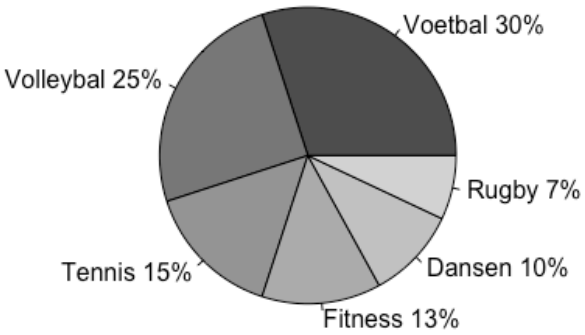
20. Gegeven zijn de scores op variabel *X*. Een onderzoeker wil de ruwe scores lineair transformeren door ze te vermenigvuldigen met 1 en er daarna 20 bij op te tellen. Wat verandert er door deze transformatie wel, en wat niet?

A. De vorm van de verdeling van scores en het gemiddelde veranderen niet, maar de standaarddeviatie wordt 20 punten groter.
B. De vorm van de verdeling van scores en de standaarddeviatie veranderen niet, maar het gemiddelde wordt 20 punten hoger.
C. De vorm van de verdeling van scores verandert niet, maar het gemiddelde en de standaarddeviatie worden 20 punten hoger.
D. De vorm van de verdeling van scores zal meer normaal verdeeld zijn, het gemiddelde en de standaarddeviatie worden 20 punten hoger.

21. In een vragenlijst staat het volgende item: 'Hoe vaak heb je afgelopen week je haar gewassen'. Het is een MC-vraag met de volgende antwoordmogelijkheden: 1 = niet, 2 = één keer, 3 = twee keer, 4 = drie keer, 5 = vier keer of vaker. Wat is het hoogst zinvolle meetniveau van deze variabele?

A. Nominaal
B. Ordinaal
C. Interval
D. Ratio

22. Van 800 random geselecteerde studenten is bekend welke sport zij (primair) beoefenen. De resultaten zijn weergegeven in onderstaand taartdiagram (pie chart). Op basis van deze gegevens, hoeveel studenten beoefenen er (ongeveer) rugby?



A. 7
B. 56
C. 80
D. 560

23. Van 500 deelnemers aan een concert van Justin Bieber is de leeftijd weergegeven in onderstaande tabel. Wat is de mediaan van de leeftijd?

Leeftijd	9	10	11	12	13	14	15	16	22
Aantal deelnemers	32	83	90	100	87	32	16	56	4

A. 11
B. 11,5
C. 12
D. 12,2

24. Er zijn drie kinderen van 1, 3 en 5 jaar in een kamer. Als er een 3-jarig kind de kamer binnenkomt, wat gebeurt er dan met het gemiddelde en de variantie?

A. Het gemiddelde blijft gelijk, maar de variantie wordt groter
B. Het gemiddelde blijft gelijk, maar de variantie wordt kleiner
C. Het gemiddelde en de variantie blijven beide gelijk
D. Het gemiddelde en de variantie worden beide kleiner

25. Een docent statistiek geeft een tentamen aan 5 studenten. Hij komt tot de volgende cijfers: 4, 6, 7, 7, 8. Wat is de variantie voor deze scores?

A. 0
B. 0.76
C. 1.40
D. 2.30

26. Wanneer kun je beter gebruiken maken van de *five-number summary* (vijf-getallen samenvatting) dan van het gemiddelde en de standaarddeviatie om de verdeling van een variabele te beschrijven?

A. Nooit, het gemiddelde en de standaarddeviatie zijn altijd beter
B. Als de verdeling van de variabele redelijk symmetrisch is
C. Als de verdeling van de variabele sterk scheef verdeeld is met sterke uitbijters
D. Als de verdeling van de variabele licht scheef verdeeld is zonder uitbijters

27. Een random variabele X heeft een gemiddelde van 10 en een standaarddeviatie van 2. De variabele X wordt vermenigvuldigd met 2 om zo een nieuwe variabele Y te maken: Y = 2X. Wat is de variantie van de nieuwe variabele Y?

A. 2
B. 4
C. 16
D. 32

28. Uit een onderzoek blijkt dat mensen die meer bier drinken, minder vaak ziek zijn. Ook blijkt dat mensen die meer bier drinken, vaker sinaasappelsap drinken. De variabelen “sinaasappelsap drinken” en “bier drinken” zijn ……. variabelen als verklaring voor minder ziek zijn.

A. Scheve
B. Normaal verdeelde
C. Verklarende
D. Verstrengelde (c*onfounding*)

29. Een groep studenten denkt dat het drinken van sinaasappelsap zorgt voor lichamelijk herstel. Om dit te testen gaan zij wekelijks naar een bejaardentehuis waar zij de bewoners bezoeken en met hen praten onder het genot van een glaasje sinaasappelsap. Na enkele maanden zijn veel van de bewoners vrolijk en gezond. Wat is de verklarende variabele in dit onderzoek?

A. Sinaasappelsap
B. De woonsituatie (bejaardentehuis)
C. De emotionele toestand van de bewoners
D. Alle bovenstaande antwoorden

30. In een grootschalig onderzoek in de V.S. worden verschillende variabelen gemeten. Welke van onderstaande variabelen is een nominale variabele?

A. De staat waarin men woont
B. De leeftijd van de respondent
C. Het aantal mensen binnen het huishouden
D. Het totale inkomen van het huishouden per jaar

31. Om te onderzoeken in hoeverre de scores op twee variabelen gelijk zijn, kan men het beste gebruiken maken van

A. De correlatie
B. Kendall's tau
C. De interkwartiel afstand
D. Het gemiddelde absolute verschil

32. Kees heeft de scores van 10 personen op een test weergegeven in een stemplot. Nu wil Kees de figuur uitbreiden door onderscheid te maken tussen mannen en vrouwen. Welke figuur kan hij hiervoor het beste gebruiken?

A. Een staafdiagram
B. Een histogram
C. Een tijdsplot
D. Een back-to-back stemplot

Antwoorden Hoofdstuk 1

1	B	De interkwartiele range is het derde kwartiel minus het eerste kwartiel, in formulevorm: IQR = $Q_3 - Q_1$
2	C	De variabele 'aantal uren studeren' verklaart (deels) het wel of niet halen van het tentamen en is daarmee een onafhankelijke variabele (ook wel verklarende variabele genoemd). Dit zegt echter niks over de verdeling van een variabele, dus er kunnen op basis van deze gegevens geen uitspraken worden gedaan over de verdeling (bijvoorbeeld of de variabele normaal of scheef verdeeld is).
3	B	Bij de eerste stemplot is de *mediaan* (middelste waarde) 73 en de *modus (*meest voorkomende waarde) 61. Bij de tweede stemplot is de mediaan 66 en de modus 55.
4	A	
5	D	De mediaan is 6. De minimum score is 4 en de maximum score is 9. Dit betekent dat de mogelijke waarden onder de mediaan variëren van 4-6 en boven de mediaan van 6-9. De spreiding boven de mediaan is dus groter dan de spreiding onder de mediaan. De five-number-summary geeft geen directe informatie over de modus.
6	A	Een boxplot geeft de mediaan, eerste en derde kwartiel, en eventuele uitbijters weer. Wanneer een variabele niet (perfect) normaal verdeeld is, is het gemiddelde niet gelijk aan de mediaan en valt het gemiddelde dus niet rechtstreeks uit een boxplot af te leiden.
7	B	
8	C	$x_{new} = a + bx$ Het vermenigvuldigen van iedere observatie met *b* (hier: 0.5) zorgt voor een vermenigvuldiging van zowel centrummaten (gemiddelde) als spreidingsmaten (variantie) met dat getal. Optellen/aftrekken van *a* bij iedere observatie zorgt voor het optellen/aftrekken van *a* bij centrummaten, maar verandert niets aan spreidingsmaten.
9	D	De standaarddeviatie wordt beinvloed door uitbijters en is dus niet resistent; een paar uitbijters kunnen de standaarddeviatie erg verhogen. De standaarddeviatie is nul wanneer er geen spreiding is, dat wil zeggen dat alle observaties dezelfde waarde hebben.
10	A	Het gemiddelde wordt getrokken naar de kant waar de staart zit, want deze wordt relatief sterk beïnvloed door extreme scores. De mediaan wordt minder beïnvloedt door extreme scores, en ligt in dit geval dus lager dan het gemiddelde.
11	B	Eerst de standaarddeviatie aanpassen: S_{nieuw}= SD * |b| geeft b = 0.5 Vervolgens alleen het gemiddelde nog aanpassen: 100 = 0.5*100 + a geeft a = 50
12	B	De interkwartiel range (IQR) = 35 – 25 = 10 1.5*IQR betekent dus dat scores beneden (25 – 1.5*10) = 10 en boven (35 + 1.5*10) = 50 als uitbijter worden aangemerkt.

13	D	Mediaan en IQR zijn relatief robuuste samenvattingsmaten
14	D	Een Q-Q plot geeft aan in hoeverre er sprake is van een normale verdeling. De lijn loopt niet mooi diagonaal, dus alleen D is goed.
15	C	Het gaat hier om een kwalitatieve (categorische) variabele. Deze kan het beste weergegeven worden met een staafdiagram. Alle andere figuren zijn alleen geschikt voor kwantitatieve (interval) variabelen.
16	D	De onafhankelijke variabele is de variabele waarmee je de afhankelijke variabele probeert te verklaren.
17	C	
18	C	Als de populatie normaal verdeeld is, zal een aselecte steekproef van 500 personen dat hoogstwaarschijnlijk ook zijn. Volgens de 65-95-99.7 vuistregel omvatten 2 standaarddeviaties links en rechts van het gemiddelde 95% van de scores. Dus 1 standaarddeviatie is ongeveer 1. Het gemiddelde van de steekproef zal rond de 7 liggen.
19	D	Hier kun je uit de alternatieven afleiden wat het moet zijn: a is het niet (want dat is het streepje in het midden van de boxplots), b is het niet (want de mediaan zou zakken als je de uitbijters weg zou laten), de IQR is het niet (want dat is de breedte van de middelste box, en die komt niet overeen met de hoogte van het vierkantje. D is juist. We zien hier een rechtsscheve verdeling, wat betekent dat het gemiddelde naar rechts (hier: 'omhoog') wordt getrokken en dus boven de mediaan ligt.
20	B	
21	B	Het gaat om categorieën, dus C en D vallen af. Omdat er wel een ordening in de categorieën zit, is ordinaal het hoogst zinvolle meetniveau.
22	B	7%, dus 0.07*800 = 56
23	C	De mediaan is het (500+1)/2 = 250,5^{ste} getal, dus het midden van getal 250 en 251. Dit getal ligt bij de leeftijd van 12 jaar.
24	B	De waarneming die erbij komt is gelijk aan het gemiddelde, dus het gemiddelde blijft gelijk. De som van de gekwadrateerde verschillen is gelijk gebleven, maar doordat je deelt door een groter getal, wordt de variantie kleiner.
25	B	Bereken eerst het gemiddelde: $\bar{x} = \frac{4+6+7+7+8}{5} = 6.4$ Bereken vervolgens de som van afwijkingen van iedere score ten opzichte van het gemiddelde en kwadraat deze som: $(x - \bar{x})^2 = (4 - 6.4)^2 + (6 - 6.4)^2 + (7 - 6.4)^2 + (7 - 6.4)^2 + (8 - 6.4)^2 = (-2.4)^2 + (-0.4)^2 + (0.6)^2 + (0.6)^2 + (1.6)^2 =$ 5.76 + 0.16 + 0.36 + 0.36 + 2.56 = 9.2 Vervolgens neem je daar de wortel van en deel je door *n-1*.

		Dus $var = \frac{1}{4}\sqrt{9.2} \approx 0.76$
26	C	
27	C	$\sigma_{a+bX} = b\ \sigma_x$, dus voor de variantie betekent dit: $\sigma^2_{a+bX} = b^2\sigma^2_x$ Dus de variantie $\sigma^2 = 2^2 * 2^2 = 4 * 4 = 16$
28	D	
29	A	
30	A	
31	D	De correlatie en Kendall's tau gaan over het verband tussen de variabelen; dit zegt niets over het verschil tussen scores. De interkwartielafstand zegt iets over de spreiding van de scores, ook hier kunnen geen uitspraken worden gedaan over het verschil/de overeenkomst tussen scores.
32	D	

Hoofdstuk 2

1. In SPSS is een regressieanalyse uitgevoerd met de variabelen educatie (aantal jaren onderwijs) en inkomen. Onderstaande tabel is de output van de regressieanalyse in SPSS. Wat zijn hier de *a* en *b* in de regressieformule $\hat{y} = a + bx$?

A. *a* = -1636.364 en *b* = 237.063
B. *a* = 237.063 en *b* = -1636.364
C. *a* = -0.606 en *b* = 1.495
D. *a* = -1636.364 en *b* = -0.606

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1636.364	2699.962		-.606	.561
	Education	237.063	158.575	.467	1.495	.173

a. Dependent Variable: Income

2. Wat probeert men te minimaliseren in een spreidingsdiagram van de regressie van Y op X?

A. De kwadratensom van horizontale afstanden van de punten tot de lijn
B. De kwadratensom van verticale afstanden van de punten tot de lijn
C. De kwadratensom van de kortste afstanden van de punten tot de lijn
D. De kwadratensom van horizontale en verticale afstanden van de punten tot de lijn

3. Gegeven is dat de correlatie tussen *X* en *Y* gelijk is aan 0.6. Verder is gegeven dat *X* een gemiddelde heeft van 3 en *Y* een gemiddelde heeft van 5. De standaarddeviatie van zowel *X* als *Y* is 1. Wat zijn *a* en *b* in de regressievergelijking $\hat{y} = a + bx$?

A. *a* = 0 en *b* = 0.6
B. *a* = 0.6 en *b* = 0
C. *a* = 0.6 en *b* = 3.2
D. *a* = 3.2 en *b* = 0.6

4. De correlaties tussen vier variabelen zijn berekend en weergegeven in onderstaande tabel. De onderzoeker wil een lineaire regressievergelijking opstellen om het tentamencijfer te voorspellen op basis van één van de andere variabelen. Uitgaande van onderstaande tabel, welke variabele is de beste voorspeller van het tentamencijfer?

A. Aantal uren gestudeerd
B. Aantal uren Netflix
C. Vorige tentamencijfer
D. Daar valt op basis van correlaties niets over te zeggen

Correlations

		Tentamencijfer	Aantal_uren_gestudeerd	Aantal_uren_Netflix	Vorige_tentamencijfer
Tentamencijfer	Pearson Correlation	1	-.277	-.952**	.533
	Sig. (2-tailed)		.438	.000	.113
	N	10	10	10	10
Aantal_uren_gestudeerd	Pearson Correlation	-.277	1	.377	.394
	Sig. (2-tailed)	.438		.283	.260
	N	10	10	10	10
Aantal_uren_Netflix	Pearson Correlation	-.952**	.377	1	-.379
	Sig. (2-tailed)	.000	.283		.280
	N	10	10	10	10
Vorige_tentamencijfer	Pearson Correlation	.533	.394	-.379	1
	Sig. (2-tailed)	.113	.260	.280	
	N	10	10	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

5. In een onderzoek naar het verband tussen overgewicht en bezoek aan de huisarts is gevonden dat mensen met overgewicht vaker naar de huisarts gaan dan mensen zonder overgewicht. Daarmee is aangetoond dat

A. Overgewicht ervoor zorgt dat mensen vaker naar de huisarts gaan
B. Mensen die overgewicht hebben minder naar de huisarts zullen gaan, wanneer ze afvallen.
C. Er een samenhang is tussen het wel of niet hebben van overgewicht en het aantal huisartsbezoeken.
D. Er onder de mensen met overgewicht veel mensen zijn die een bezoek aan de huisarts brengen.

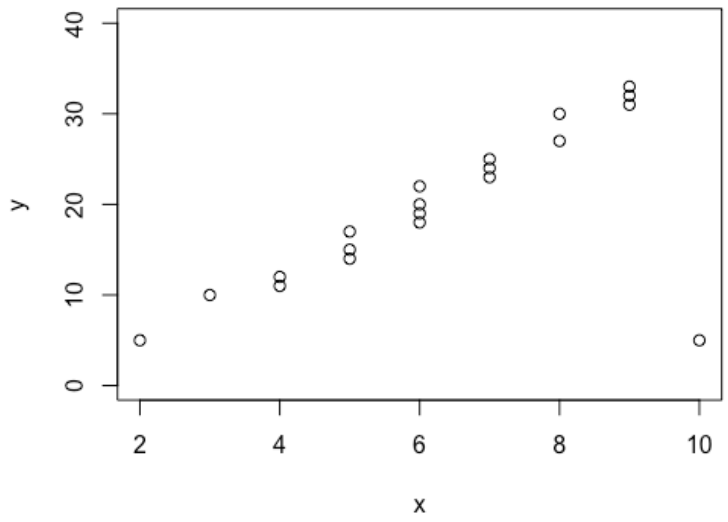
6. Gegeven zijn de scores van 100 proefpersonen. We weten dat de variantie van *X* gelijk is aan 4 en dat de variantie van *Y* gelijk is aan 9. De covariantie tussen *X* en *Y* is gelijk aan 3. Wat is dan de correlatie tussen *X* en *Y*?

A. 0.08
B. 0.25
C. 0.50
D. 0.75

7. In een onderzoek naar het verband tussen gebit en geheugen (Algemeen Dagblad, 2004) is gevonden dat mensen die hun eigen gebit nog hadden een beter geheugen hadden dan mensen met een kunstgebit. De onderzoekers concluderen dat 'tanden en kiezen uiterst belangrijk zijn voor ons geheugen'. Een criticus beweert echter dat het gevonden verband eenvoudig te verklaren is via *lurking variables* (derde variabelen). Welke van onderstaande variabele(n) kan hier de rol van een derde variabele spelen?

A. Het al dan niet hebben van een kunstgebit
B. De leeftijd
C. Het geheugen
D. Alle drie de bovenstaande variabelen

8. In onderstaande figuur zijn de scores van 20 personen op variabelen *X* en *Y* weergegeven. Van de 20 personen, valt één persoon nogal op. Vormen de scores van deze persoon een invloedrijk punt?



A. Ja, want het weglaten van deze persoon zal de correlatie tussen *X* en *Y* aanzienlijk veranderen.
B. Ja, want de score van deze persoon op variabele *Y* kan duidelijk als *outlier* (uitbijter) worden opgevat.
C. Nee, want het weglaten van deze persoon zal de correlatie tussen *X* en *Y* niet veranderen.
D. Nee, want de score van deze persoon op variabelen *X* en *Y* kunnen duidelijk niet als *outlier* (uitbijter) worden opgevat.

9. De correlatie tussen variabelen *X* en *Y* blijkt precies 1.0 te zijn. Wat mag je concluderen?
A. Het gemiddelde absolute verschil zal 0 zijn
B. De helling van de regressievergelijking zal gelijk zijn aan 0
C. De scores op X zijn gelijk aan de scores op Y
D. De scores op Y zijn een lineaire transformatie van de scores op X

10. Gegeven zijn twee variabelen *X* en *Y*. Om *Y* te voorspellen uit *X* is de volgende regressievergelijking opgesteld: $\hat{y} = -9 + 3.2X$. De correlatie tussen *X* en *Y* is 1.0. Als je weet dat iemand een score heeft van -9 op Y, wat kun je dan zeggen over het residu $y - \hat{y}$?
A. Het residu zal positief zijn
B. Het residu zal negatief zijn
C. Het residu zal nul zijn
D. Daar kan op basis van deze gegevens geen uitspraak over worden gedaan

11. Gegeven is dat de correlatie tussen *X* en *Y* gelijk is aan -0.40. Beide variabelen hebben een gemiddelde van 30. De standaarddeviatie van *X* is 6. De standaarddeviatie van *Y* is 3. Wat is het intercept in de regressievergelijking van *Y* op *X*?
A. 6
B. 24

C. 36
D. 54

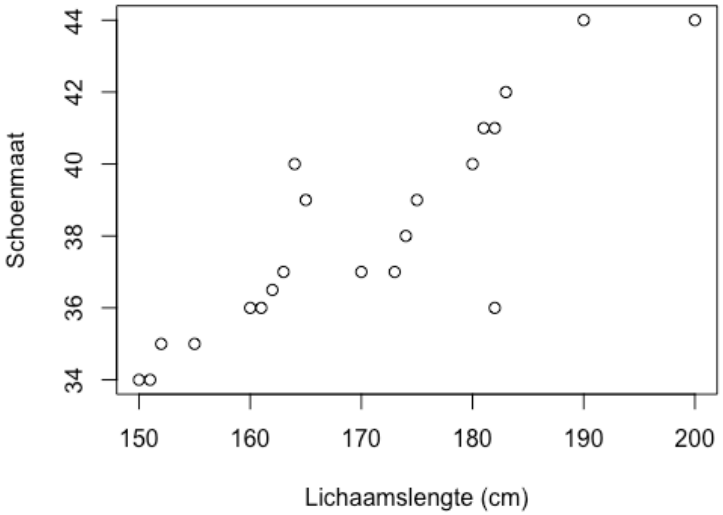
12. Gegeven is dat de correlatie tussen *X* en *Y* gelijk is aan 0. Hieronder staan 4 conclusies die zijn getrokken op basis van dit gegeven. Welke conclusie is onjuist?

A. Er is geen lineaire samenhang tussen *X* en *Y*
B. De scores op *X* en *Y* zijn volledig identiek aan elkaar
C. De regressievergelijking is een horizontale lijn (helling = 0)
D. Er is 0% verklaarde variantie zijn bij een lineaire regressie

13. In welke situatie is er sprake van Simpson's paradox?

A. Ziekenhuis X heeft een lager sterftecijfer bij terminale patiënten, en ziekenhuis Y heeft een lager sterftecijfer bij niet-terminale patiënten. Als we het al dan niet terminaal zijn van patiënten buiten beschouwing laten heeft ziekenhuis X een lager sterftecijfer
B. Ziekenhuis X heeft een lager sterftecijfer bij terminale patiënten, en ziekenhuis Y heeft een lager sterftecijfer bij niet-terminale patiënten. Als we het al dan niet terminaal zijn van patiënten buiten beschouwing laten heeft ziekenhuis Y een lager sterftecijfer
C. Ziekenhuis X heeft een lager sterftecijfer bij terminale patiënten, en ziekenhuis X heeft een lager sterftecijfer bij niet-terminale patiënten. Als we het al dan niet terminaal zijn van patiënten buiten beschouwing laten heeft ziekenhuis X een lager sterftecijfer
D. Ziekenhuis X heeft een lager sterftecijfer bij terminale patiënten, en ziekenhuis X heeft een lager sterftecijfer bij niet-terminale patiënten. Als we het al dan niet terminaal zijn van patiënten buiten beschouwing laten heeft ziekenhuis Y een lager sterftecijfer

14. Wat is onderstaand scatterplot een redelijke schatting van de correlatie tussen lichaamslengte in centimeters en schoenmaat?



A. -0.70
B. -0.10
C. 0.10
D. 0.70

15. Er is een lineaire regressievergelijking opgesteld: $y = 10 + 0.8x$, waarbij *y* de eindscore is en *x* de deelscore is. Stel dat Marleen 80 scoort op de deelscore, wat is dan haar voorspelde (predicted) eindscore?

A. 64
B. 72
C. 74
D. 80

16. Iemand vraagt zich af of vrouwen daten met partners met gelijke lichaamslengte. In onderstaande tabel staat de data weergegeven van de lichaamslengte in inches (1 inch ≈ 2.5 cm) van zes vrouwen en hun date.

Lengte vrouw	64	65	65	66	66	68
Lengte date	68	69	69	70	72	73

Welke van de onderstaande uitspraken is juist?

A. Elke lichaamslengte boven de 66 inches moet beschouwd worden als uitbijter.
B. Er is een sterke positieve samenhang tussen de lichaamslengte van de vrouwen en hun date
C. Er is een sterke negatieve samenhang tussen de lichaamslengte van de vrouwen en hun date
D. Als de lichaamslengte van de vrouwen en hun data uitgedrukt zou zijn in centimeters, dan zou de correlatie 2,5 maal groter zijn

17. In een onderzoek naar het verband tussen geslacht en inkomen blijkt dat de correlatie tussen deze twee variabelen gelijk is aan *r* = -0.61. Welke van onderstaande uitspraken is juist?

A. Vrouwen verdienen gemiddeld meer dan mannen
B. Mannen verdienen gemiddeld meer dan vrouwen
C. Er is een rekenfout gemaakt, de correlatie moet positief zijn
D. De meting is zinloos; *r* kan alleen bepaald worden voor twee kwantitatieve variabelen

18. Veel middelbare scholieren in de VS maten de SAT-test en/of de ACT-test als toelatingstest voor vervolgonderwijs. Er zijn data verzameld van 60 scholieren die zowel de SAT-test als de ACT-test hebben gemaakt.

- De SAT-test had een gemiddelde van 888 met een standaarddeviatie van 180
- De ACT-test had een gemiddelde van 25 met een standaarddeviatie van 5
- De correlatie tussen de SAT-test en ACT-test is 0.851

Een onderzoeker wil de SAT voorspellen uit de ACT met behulp van een lineaire regressievergelijking. Wat is de kleinste kwadraten regressielijn $y = ax + b$ bij deze data?

A. $y = 122.10 + 30.636x$

B. $y = 30.636 + 122.10x$
C. $y = 0.024 + 3.725x$
D. $y = 3.725 + 0.024x$

19. Er wordt een kleinste kwadraten regressielijn geschat voor een variabele. Een van de data-punten heeft een positief residu. Welke van de onderstaande uitspraken is juist?

A. De correlatie tussen alle voorspelde en geobserveerde datapunten is positief
B. Dit data-punt ligt boven de regressielijn
C. Dit data-punt moet een invloedrijk punt zijn
D. Dit data-punt ligt aan de rechterkant van het spreidingsdiagram

Antwoorden Hoofdstuk 2

1	A	*a* is intercept, *b* is de slope (helling).
2	B	
3	D	$b = r_{xy}\frac{s_y}{s_x} = 0.6\ \frac{1}{1}$ dus $b = 0.6$ $a = \bar{y} - b * \bar{x} = 5 - 0.6 * 3 = 3.2$ dus $a = 3.2$
4	B	$r^2 = (-0.952)^2 = 0.906$. Dus het aantal uren Netflix kijken verklaart ongeveer 90% van de variantie op het tentamencijfer.
5	C	A en B impliceren een oorzakelijk verband. D is onjuist, omdat het zo zou kunnen zijn dat er onder de mensen met overgewicht weinig mensen zijn die de huisarts bezoeken, maar wel meer dan mensen zonder overgewicht. Het zegt dus iets over het relatieve aantal huisartsbezoeken van mensen met overgewicht ten opzichte van mensen zonder overgewicht, niet over het absolute aantal huisartsbezoeken.
6	C	$r_{xy} = \frac{cov(x,y)}{S_x S_y} = \frac{3}{\sqrt{4}*\sqrt{9}} = \frac{3}{2*3} = \frac{3}{6} = 0.5$
7	B	Een derde variabele is een variabele –anders dan de verklarende of veroorzakende variabele- die van invloed is/kan zijn op de relatie tussen variabelen in een studie.
8	A	
9	D	Correlatie geeft in hoeverre er sprake is van het op één lijn liggen van de scores: een correlatie van 1 duidt er dus op dat ze precies op 1 lijn liggen. Dit betekent niet per se dat de scores gelijk zijn, of dat de helling 1 is. Als de scores niet per se gelijk zijn hoeft het verschil dus ook niet 0 te zijn.
10	C	Een correlatie van 1 betekent dat alle punten perfect op één lijn liggen (zie ook vorige vraag). Een gevolg hiervan is dat alle residuen 0 zijn.
11	C	$b = \frac{S_y}{S_x} * r_{xy} = \frac{3}{6} * -.40 = -0.20$ $a = \bar{y} - b * \bar{x} = 30 - 0.20 * 30 = 30 - -6 = 30 + 6 = 36$
12	B	Een correlatie van 0 betekent dat er geen lineair verband is. Dit betekent dat A en C juist zijn. Het percentage verklaarde variantie is r^2 en is dus ook 0. B is onjuist.
13	D	Letterlijk besproken in College. Zie ook pagina 143-145 in het boek voor een gedetailleerde uitleg met een ander voorbeeld. Kern van Simpson's paradox: een verband dat er oorspronkelijk lijkt te zijn, draait om als je een derde variabele toevoegt.
14	D	De regressielijn loopt omhoog; er is dus sprake van een positieve correlatie. Daarnaast is te zien dat er een redelijke samenhang is tussen

		lichaamslengte en schoenmaat; D benadert dit verband het beste.
15	C	y = 10 + 0.8*80 = 74
16	B	
17	D	
18	A	$b = r * \frac{S_{SAT}}{S_{CAT}} = 0.851 * \frac{180}{5} = 30.636$ $a = SAT - b * ACT = 888 - 30.636 * 25 = 122.1$
19	B	

Hoofdstuk 3

1. Wat is een voorbeeld van een *matched-pairs design* met twee condities?
   A. Elke proefpersoon wordt verbonden aan een vergelijkbare proefpersoon. Deze twee proefpersonen worden random aan een van de condities toegewezen en vergeleken.
   B. Elke proefpersoon wordt achtereenvolgens toegewezen aan beide condities. De volgorde van de condities wordt random gekozen per proefpersoon.
   C. Geen van beide
   D. Beide

2. Een random steekproef is een steekproef waarbij
   A. De proefpersonen uit random uit de populatie worden getrokken
   B. De condities at random worden toegewezen aan de proefpersonen
   C. De condities at random worden geselecteerd
   D. De condities in een random volgorde worden toegewezen aan de proefpersonen

3. Welke van de volgende uitspraken over experimenteel onderzoek is juist?
   A. De onafhankelijke variabele wordt gemanipuleerd door de onderzoeker
   B. Het is bij een experiment mogelijk een causaal verband te onderzoeken

   E. Alleen bewering I is waar
   F. Alleen bewering II is waar
   G. Beide beweringen zijn waar
   H. Beide beweringen zijn niet waar

4. Een onderzoeker wil een studie doen naar de relatie tussen inkomen en opleidingsniveau. Hij wil bij het verzamelen van zijn gegevens rekening houden met de verhouding tussen mannen en vrouwen (die in de populatie 50% om 50 % is), en met de verhouding in sociaaleconomische status (SES). SES is onderverdeeld in drie categorieën: laag, gemiddeld en hoog, die in de populatie respectievelijk bij 30%, 60, en 10% voorkomen. Om deze percentuele verhoudingen exact terug te vinden in zijn steekproef categoriseert hij de populatie volgens geslacht en SES, en vervolgens trekt hij uit iedere groep een bepaald aantal mensen (in de verhouding zoals ze voorkomen in de populatie). Wat voor type steekproef beschrijft deze manier van steekproeftrekking het best?

   A. Convenient sample
   B. Stratified sample
   C. Multistage sample
   D. Paired sample

5. Anneloes is flink verkouden. Haar huisgenoot slikt elke dag een knoflooktablet en is al twee jaar lang niet verkouden geweest. De tante van Anneloes heeft een kennis die ook dagelijks een knoflooktablet inneemt en ook al meer dan een jaar niet verkouden is geweest. Op basis van deze gegevens besluit Anneloes om knoflooktabletten te gaan

innemen zodra haar verkoudheid voorbij is. Op welk onderzoek is Anneloes haar beslissing gebaseerd?

A. Anekdotisch bewijs
B. Een observationeel onderzoek gebaseerd op beschikbare data
C. Een observationeel onderzoek gebaseerd op een steekproef
D. Een experiment

6. De samenhang tussen cola drinken en gewichtstoename is onderzocht. De studie bestond uit 25 deelnemers, ingedeeld in twee groepen. De eerste groep deelnemers volgde een cola-vrij dieet. De twee groep volgde een cola-rijk dieet. Na 8 weken is de gewichtstoename van iedere deelnemer gemeten. Dit onderzoek is een voorbeeld van een

A. Observationeel onderzoek
B. Survey
C. Matched-pairs experiment
D. Experiment, maar niet een dubbelblind experiment

7. Geertje wil prijsverschillen van koffiemelk onderzoeken bij Albert Heijn, Jumbo en De Spar. Hoe kan Geertje het beste de producten kiezen om bias zoveel mogelijk te voorkomen?

A. Kies veel gekochte soorten koffiemelk
B. Kies koffiemelk van bekende merken
C. Kies zowel veel gekochte als merkproducten
D. Selecteer random een aantal beschikbare producten

8. Bij een onderzoek naar Ritalin worden 100 vrijwilligers eerst ingedeeld naar geslacht. Daarna krijgt de helft van de mannen (random toegewezen) Ritalin, en de andere helft een placebo. Ook bij de vrouwen krijgt de helft (random toegewezen) Ritalin en de andere helft een placebo. Dit is een voorbeeld van een

A. Replicatie
B. Matched-pairs design
C. Verstrengeling, want het effect van geslacht raakt verstrengeld met het effect van het medicijn
D. Block-design

Antwoorden Hoofdstuk 3

1	D	Een matched pairs design kan zowel betrekking hebben op de toewijzing/volgorde van proefpersonen aan beide condities, als het toewijzen van vergelijkbare proefpersonen aan verschillende condities.
2	A	
3	C	
4	B	De populatie wordt opgedeeld in ‘strata’. Vervolgens wordt uit iedere stratum een steekproef genomen. Op die manier worden de verhoudingen in de populatie behouden.
5	A	
6	D	
7	D	
8	D	

Hoofdstuk 4

1. Gegeven zijn de scores op variabele *X* met een gemiddelde van 10 en een standaarddeviatie van 2. Op grond hiervan kunnen de scores op variabele *Y* berekend worden met $Y = 10 - 2X$. De standaarddeviatie van Y is

A. 2
B. 4
C. 16
D. 32

2. Gegeven zijn twee gebeurtenissen *A* en *B.* Gegeven is dat *P(B)* = 0.6, *P(A* en *B)* = 0.3 en *P(A* of *B)* = 1.0. Wat is dan de kans op gebeurtenis *A*, oftewel *P(A)*?

A. 0.1
B. 0.3
C. 0.6
D. 0.7

3. Gegeven zijn twee gebeurtenissen *A* en *B*. Gegeven is dat *P*(A) = 0.3 en *P(B)* = 0.5 en *P(B|A)* = 0.8. Wat is dan de kans op *P(A* en *B*)?

A. 0.15
B. 0.24
C. 0.40
D. 0.48

4. Er wordt twee keer geworpen met een eerlijke dobbelsteen. Hoe groot is de kans dat de som van beide worpen gelijk is aan 12?

A. 1/36
B. 2/36
C. 4/36
D. 1/12

5. Van een groep ouderen zijn de volgende gegevens bekend over de woonsituatie en eenzaamheid (zie onderstaande tabel).

	Eenzaam		Totaal
	Wel	Niet	
In bejaardentehuis	40	30	70
Zelfstandig wonend	10	20	30
Totaal	50	50	100

Wat is de kans dat een oudere, waarvan bekend is dat hij of zij in het bejaardentehuis woont, eenzaam is?

A. 40/70
B. 40/100
C. 50/100
D. 70/100

6. Mensen die psychotisch zijn, zijn vaak ook depressief. Om dit te onderzoeken zijn gegevens verzameld van 100 patiënten. Gegeven is dat 30% van de patiënten psychotisch is. Van de patiënten die psychotisch zijn, is 80% depressief. Van de patiënten die niet psychotisch zijn, is slechts 20% depressief. Hoeveel patiënten uit deze steekproef zijn psychotisch en depressief?

A. 20
B. 24
C. 30
D. 80

7. Als gebeurtenissen A en B afhankelijk zijn, dan geldt:

A. P(A | B) = 0
B. P (A en B) = 0
C. Zowel A als B
D. Geen van bovenstaande antwoorden is juist

8. Gegeven is dat 25% van de mensen een vitaminetekort heeft. Verder is bekend dat van alle mensen met een vitaminetekort, 80% hier ook daadwerkelijk positief op test. Bij mensen die geen vitaminetekort hebben, blijkt 10% toch een positief testuitslag te hebben. Wat is de kans dat iemand die een positieve uitslag krijgt ook daadwerkelijk een vitaminetekort heeft?

A. 20%
B. 73%
C. 80%
D. 90%

9. Gegeven is de onderstaande kansverdeling op variabele *X*. Het gemiddelde van *X* is 2.5. Wat is de verwachte standaarddeviatie van deze variabele?

X	1	2	3	4
P	.30	.20	.20	.30

A. 1.20
B. 1.45
C. 1.80
D. 2.00

10. Gegeven is: P(A) = 0.40 en P(B) = 0.30. Verder is bekend dat A en B onafhankelijk zijn. Wat is de kans op A gegeven B?

A. 0.12
B. 0.30
C. 0.40
D. Dat is niet te bepalen zonder meer gegevens

11. Bij het spelletje moet je drinken wanneer je 1 gooit. Iemand doet 3 rondes mee met dit spelletje, en werpt met dezelfde eerlijke dobbelsteen. Wat is de kans dat deze persoon precies 1 keer moet drinken?

A. $1 * (\frac{1}{6}^{1} * \frac{5}{6}^{2})$
B. $3 * (\frac{1}{6}^{1} * \frac{5}{6}^{2})$
C. $\binom{3}{2}$
D. Dit is niet te bepalen aan de hand van deze gegevens.

12. Stel dat A en B twee onafhankelijke gebeurtenissen zijn. Gegeven is dat P(A) = 0.5 en P(B) = 0.2. Wat is de kans dat A niet gebeurt en dat B niet gebeurt?

A. 0.1
B. 0.3
C. 0.4
D. 0.7

13. Wanneer je twee keer gooit met een eerlijke dobbelsteen, hoe groot is dan de kans dat je beide keren hetzelfde getal gooit?

A. 1/6
B. 1/12
C. 1/18
D. 1/36

14. Gegeven is de onderstaande kansverdeling van X, waarbij X het aantal cursussen is dat een voltijdstudent heeft gevolgd deze periode.

X	1	2	3	4
P	.20	.30	.20	.30

Wat is het gemiddeld aantal gevolgde cursussen door voltijdstudenten deze periode?

A. 0.65
B. 2
C. 2.6
D. 3

15. En wat is de standaarddeviatie van de variabele X, zoals weergegeven bij vraag 14?

A. 0.32
B. 0.64
C. 1.04
D. 1.10

16. Hans wordt regelmatig ingehuurd om bepaalde computerproblemen op te lossen, zo ook het debuggen van virussen. Recent zijn er twee virussen in omloop: virus Dummy en virus Smarty. De volgende gegevens zijn bekend:

- 65% van de klanten heeft problemen met virus Dummy en 35% heeft problemen met virus Smarty

- Als de computer besmet is met Dummy, dan is er 80% kans dat Hans de problemen kan oplossen
- Als de computer besmet is met Smarty, dan is er 30% kans dat Hans de problemen kan oplossen

Als er random een computer geselecteerd wordt, waarvan we weten dat Hans de problemen heeft opgelost, wat is dan de kans dat deze computer besmet was met Dummy?

A. 0.52
B. 0.53
C. 0.63
D. 0.83

17. Gegeven zijn twee disjuncte gebeurtenissen A en B. De kans op gebeurtenis A is 0.2 De kans op gebeurtenis B is 0.8. Wat is P(A of B)?

A. 0.6
B. 0.8
C. 1.0
D. Dat is niet te bepalen zonder meer gegevens

Antwoorden Hoofdstuk 4

1	C	$var(Y) = (-2)^2 * var(X) = (-2)^2 * (2)^2 = 4 * 4 = 16$ $sd(Y) = \sqrt{var(Y)} = \sqrt{16} = 4$
2	D	$P(A\ of\ B) = P(A) + P(B) - P(A\ en\ B)$. Deze formule invullen geeft: $1.0 = x + 0.6 - 0.3$. *Dus* $x = 1.0 - 0.6 + 0.3 = 0.7$
3	B	$P(A\ en\ B) = P(B\|A) * P(A) = 0.24$
4	A	De som van twee worpen is alleen gelijk aan 12 als beide keren een 6 wordt gegooid. Dus: $\frac{1}{6} * \frac{1}{6} = \frac{1}{36}$
5	A	Er wordt gevraagd om een conditionele kans (= gegeven woonsituatie in bejaardentehuis).
6	B	30% is psychotisch, dus 30/100 * 100 = 30 patiënten zijn psychotisch. Van die 30 personen, is 80% depressief. Dus: 80/100 * 30 = 24 patiënten zijn psychotisch depressief. Tip: teken een boomdiagram.
7	D	Onafhankelijk betekent dat de ene gebeurtenis geen invloed of voorspellende waarde heeft op de andere gebeurtenis. Als twee gebeurtenissen onafhankelijk zijn, zegt A niks over de kans op B: P(B|A) = P(B) en A en B kunnen gerust samen optreden.
8	B	Teken een boomdiagram. Uitgaande van 1000 personen hebben in totaal 275 mensen een positieve testuitslag, waarvan 200 mensen ook daadwerkelijk een vitaminetekort hebben (immers: 250*0.8 = 200). Dat komt overeen met 73% (want: 200/275 * 100 $\approx 73\%$).
9	A	$Var = 0.30 * (1 - 2.5)^2 + 0.20 * (2 - 2.5)^2 + 0.20 * (3 - 2.5)^2$+$0.30 * (4 - 2.5)^2$ = 1.45 $SD = \sqrt{1.45} \approx 1.20$
10	C	Gevraagd wordt wat de kans op A gegeven B is, oftewel: P(A|B). Als A en B onafhankelijk zijn, voorspelt B niks over A. De kans op A wordt dus niet niet beïnvloedt door de kans op B, en dus geldt: P(A|B) = P(A).
11	B	De kans op precies 1 keer drinken betekent dat je of de eerste, of de tweede, of de derde keer 1 gooit. Dit betekent 3 boven 1 (=3) maal de kans op ieder van de drie mogelijkheden (dat is 1/6 * 5/6.
12	C	P(A niet en B niet) = P(A niet) * P(B niet) = (1 – 0.5) * (1 – 0.2*) = 0.5 * 0.8 = 0.4
13	A	De kans op een bepaald getal = 1/6 De kans om dat getal beide keren te gooien = 1/6 * 1/6 = 1/36 Dit kan voor alle 6 de getallen, dus 1/36 * 6 = 6/36 ofwel 1/6
14	C	$\mu = 1 * 0.2 + 2 * 0.3 + 3 * 0.2 + 4 * 0.3 = 2.65$
15	A	Gemiddelde = 2.6 (zie vraag 14). Variantie = $(0.20 * 1\text{-}2.6)^2 + (0.30 * 2\text{-}2.6)^2 + (0.20 * 3\text{-}2.6)^2 + (0.30 * 4\text{-}2.6)^2$ = $(-0.32)^2 + (-0.18)^2 + (0.08)^2 + (0.42)^2 = 0.3176 \approx 0.32$
16	D	Maak een boomdiagram 100*0.65 = 65 > 65*0.8 = 52 (computers met Dummy, gemaakt door Hans) 100*0.35 = 35 > 35*0.3 = 10.5 (computers met Smarty, gemaakt door Hans). Dus, 52 / (52+10.5) = 0.8333 ≈ 0.84
17	C	Disjunct betekent dat P(A of B) = 1.0

Hoofdstuk 5

1. De scores op de Cito-toets zijn bij benadering normaal verdeeld met een gemiddelde van 535 en een standaarddeviatie van 5. Welk percentage van de leerlingen heeft naar schatting hoger gescoord dan 545?

A. 1%
B. 2.5%
C. 5%
D. 10%

2. Gegeven is dat de scores op de variabele inslaaptijd voor kinderen normaal verdeeld zijn met gemiddelde van 1500 seconden en een standaarddeviatie van 300 seconden. Wat is de proportie van kinderen die in meer dan 1000 seconden inslaapt?

A. 0.0475
B. 0.1423
C. 0.8577
D. 0.9525

3. Welke van onderstaande beweringen over *sampling variability* (steekproeffluctuatie) is/zijn juist?

I. De steekproeffluctuatie kan worden verkleind door de steekproef te vergroten.
II. De steekproeffluctuatie is de mate van spreiding van een statistic wanneer de statistic bij vele random steekproeven uit dezelfde populatie wordt berekend.

A. Alleen bewering I is waar
B. Alleen bewering II is waar
C. Beide beweringen zijn waar
D. Beide beweringen zijn niet waar

4. De scores op een test voor het ontwikkelingsniveau van peuters zijn normaal verdeeld met een gemiddelde van 100 en een standaarddeviatie van 10. Wat is de kans dat een willekeurige peuter een score van 115 of hoger heeft op deze test?

A. .0068
B. .4404
C. .5596
D. .9332

Gebruik de volgende gegevens voor vraag 5 en 6: De populatie Nederlandse psychologiestudenten is vrij scheef verdeeld voor geslacht: slechts 20% is man en 80% is vrouw. Gekeken wordt naar het aantal mannen in een willekeurige steekproef van psychologiestudenten (dus waarvoor geldt: $p = 0.20$).

5. Wat is de kans op minder dan *2* mannelijke studenten in een willekeurige steekproef van 8?

A. .1678 + .3355
B. .1678 + .3355 + .2936
C. 1 – (.1678 + .3355)
D. Daar kan op basis van deze gegevens geen uitspraak over worden gedaan

6. Wat is de kans op minstens 30 mannelijke studenten in een willekeurige steekpoef van 120 studenten? Gebruik hiervoor de normaal benadering van de binomiale verdeling.

A. $P(Z > 1.15)$
B. $P(Z > 1.26)$
C. $P(Z > 1.37)$
D. $P(Z > 1.48)$

7. Gegeven zijn de scores op een Cito-toets. Bekend is dat de scores in de populatie normaal verdeeld zijn met een gemiddelde van 100. In een aselecte steekproef van 25 mensen uit deze populatie is het gemiddelde 105. De standaarddeviatie in de steekproef is 3. Welke van de volgende uitspraken is juist?

A. 100 is een parameter, 25 is een statistic
B. 100 is een parameter, 105 is een statistic
C. 25 is een parameter, 3 is een statistic
D. 25 is een parameter, 105 is een statistic

8. Met een *unbiased* (zuivere) statistic wordt bedoeld dat bij een groot aantal vergelijkbare steekproeven uit dezelfde populatie, van dezelfde steekproefgrootte *n* …

A. De statistics allemaal dicht bij elkaar liggen
B. Het gemiddelde van de statistics gelijk is aan de parameter
C. De spreiding van de statistics nul is
D. Het gemiddelde van de statistics nul is

9. Wat is $P(-0.55 < Z < 1.21)$ als we gebruik maken van tabel A voor standard normaal verdelingen?

A. 0.2912
B. 0.5957
C. 0.7088
D. 0.8869

10. De scores van leerlingen op de American College Test (ACT) zijn in de populatie normaal verdeeld met een gemiddelde van 18 en een standaarddeviatie van 6. Op een bepaalde school maken 50 leerlingen de ACT. Veronderstel dat deze 50 scores dezelfde verdeling volgen als in de populatie. Wat is de steekproevenverdeling van de gemiddelde score op de ACT voor steekproeven van 50 leerlingen?

A. Ongeveer een normale verdeling, maar de benadering is slecht
B. Een exact normale verdeling
C. Een rechtsscheve verdeling
D. Een linksscheve verdeling

11. Het geboortegewicht van voldragen baby's is normaal verdeeld met een gemiddelde van 7 pond en een standaarddeviatie van 0.8 pond. Wat is de kans dat het gemiddelde gewicht van een aselect gekozen voldragen baby meer dan 7.6 pond is?

A. 0.23
B. 0.75
C. 0.77
D. Dat is niet te bepalen zonder meer gegevens

12. X is binomiaal verdeeld met parameters n = 10 en p = 0.7. Wat is het gemiddelde aantal successen, en wat is de standaarddeviatie?

A. μ = 1.45, σ = 7
B. μ = 1.45, σ = 2.1
C. μ = 7, σ = 2.1
D. μ = 7, σ = 1.45

13. Gegeven is dat 30% van de huwelijken in Nederland binnen 15 jaar eindigt in een scheiding. Een groot onderzoek heeft gedurende de laatste 15 jaar honderden huwelijken in Nederland gevolgd. Stel dat 100 van deze huwelijken aselect geselecteerd worden, wat is dan de kans dat minder dan 20 van deze huwelijken eindigen in een scheiding?

A. .011
B. .110
C. .890
D. .989

14. Gegeven is dat variabele X in de populatie sterk linksscheef verdeeld is. Hoe ziet dan de steekproevenverdeling van X eruit voor steekproeven van grote n = 100 uit deze populatie?

A. Sterk linksscheef verdeeld, in overeenstemming met de verdeling in de populatie
B. Meer normaal verdeeld dan in de populatie
C. Exact normaal verdeeld
D. Daar is op basis van deze gegevens geen uitspraak over te doen

15. Een voorwaarde voor de binomiale verdeling is dat alle observaties …. zijn

A. Onafhankelijk
B. Random
C. Afhankelijk
D. Positief

16. Er wordt een enkelvoudige aselecte steekproef getrokken uit een grote populatie. Het percentage respondenten in de steekproef met een bepaald kenmerk wordt bepaald. Wat is de beste beschrijving van dit percentage?

A. Het is een parameter
B. Het is een statistic
C. Het is een lurking variable
D. Geen van bovenstaande uitspraken is juist

Antwoorden Hoofdstuk 5

1	B	545 – 535 = 10. Dat zijn twee standaarddeviaties boven het gemiddelde. Twee standaarddeviaties links en rechts van het gemiddelde komt overeen met 95% van alle scores. Dan blijft er dus 5% over: 2.5% links (< 525) en 2.5% rechts (> 545). Tip: teken een normaalverdeling met lijnen voor het gemiddelde en de kritieke waarden.
2	D	Er wordt gevraagd hoeveel kinderen in meer dan 1000 seconden inslapen. $Z > \frac{x-\mu}{\sigma} = \frac{1000-1500}{300} = \frac{-500}{300} = -1.67$ Z = -1.67 opzoeken in Tabel A levert *p* = .0475 op. Dit is de linker overschrijdingskans. Omdat gevraagd wordt naar de proportie kinderen die meer dan 1000 seconden nodig heeft om in te slapen, doe je 1 – 0.0475 = 0.9525. Dus, 95% van de kinderen heeft een inslaaptijd van 1000 seconden of meer.
3	C	
4	A	$Z > \frac{x-\mu}{\sigma} = \frac{115-100}{10} = \frac{15}{10} = 1.5$ Z = 1.5 opzoeken in tabel A levert een linkeroverschrijdingskans op van *p* = .9932. Gevraagd wordt de rechteroverschrijdingskans (115 of meer), dus het antwoord is 1 – 0.9932 = 0.0068.
5	A	Opzoeken in Tabel C: P(*X* < 2 | p = 0.20, *n* = 8) = P(*X* = 0 | p = 0.20, *n* = 8) + P(*X* = 1 | p = 0.20, *n* = 8) .
6	B	Eerst bereken je het gemiddelde en de standaarddeviatie: $\bar{x} = 120 * 0.20 = 24$ $SD = \sqrt{n * p * (1-p)} = \sqrt{120 * 0.20 * 0.80} \approx 4.38$ Gebruik bij een normaal benadering van een binomiale verdeling altijd de continuïteitscorrectie. Dat betekent in dit geval dat je de grens voor 29.5 gebruikt in plaats van 30. Dit geeft: $P(X \geq 30 \mid p = 0.20, n = 120) = P\left(Z > \frac{29.5-24}{4.38}\right) = P(Z > 1.26)$
7	B	20 is de waarde die je wilt weten van de populatie (het gemiddelde in de populatie), 18 is de waarde berekend op basis van de steekproef (het gemiddelde in de steekproef). Onthoud: <u>p</u>opulatie > <u>p</u>arameter en <u>s</u>teekproef > <u>s</u>tatistic (pp – ss).
8	B	*Unbiased* betekent dat er geen structurele vertekening is. Dit betekent dat een individuele steekproef wellicht afwijkt van de populatieparameter, maar dat deze gemiddeld genomen gelijk zijn aan de parameter.
9	B	P(-0.55 < Z < 1.21) = P(Z < 1.21) – P(Z < -0.55) = 0.8869 – 0.2912 = 0.5957
10	B	
11	B	$Z > \frac{x-\mu}{\sigma} = \frac{7.6-7}{0.8} = \frac{0.6}{0.8} = 0.75$ geeft .7734 We willen de rechteroverschrijdingskans weten, dus P = 1 - .7734 = 0.2266
12	D	$\mu = np = 10 * 0.7 = 7$ $\sigma = \sqrt{(np(1-p)} = \sqrt{2.1} \approx 1.45$

13	A	Gebruik de normaal benadering van de binomiale verdeling met continuïteitscorrectie. $\mu = np = 0.30 * 100 = 30$ $\sigma = \sqrt{(np(1-p)} = \sqrt{30(0.70)} = \sqrt{21} \approx 4{,}58$ $P\ Z < \frac{19.5-30}{4.58} \approx -2.29$, opzoeken in tabel A geeft P < .0110
14	D	Centrale limiet theorie (zie Hoofdstuk 5, pagina 300 in het boek van Moore, McCabe & Graig).
15	A	
16	B	

Hoofdstuk 6

1. Het aantal jaren opleidingsniveau is gemeten bij een random steekproef uit de populatie van Nederlandse mannen. Vervolgens is een 95% betrouwbaarheidsinterval opgesteld voor het eerste kwartiel. Dit 95% betrouwbaarheidsinterval bevat

A. De laagste 25% van de scores op 'aantal jaren opleidingsniveau' in de steekproef
B. De laagste 25% van de scores op 'aantal jaren opleidingsniveau' in de populatie
C. Met 95% zekerheid de waarde van het eerste kwartiel in de steekproef
D. Met 95% zekerheid de waarde van het eerste kwartiel in de populatie

2. Stel we hebben het gemiddelde berekend van scores op een variabele *X* voor een random steekproef van 100 studenten uit de populatie van studenten in Groningen en we stellen een 95% betrouwbaarheidsinterval op. Dan is dit 95% betrouwbaarheidsinterval het interval waarin

A. 95% van de gevonden gemiddelden uit de steekproef liggen
B. 95% van de gevonden gemiddelden uit de populatie liggen
C. Met 95% zekerheid de steekproefwaarde van het gemiddelde van *X* ligt
D. Met 95% zekerheid de populatiewaarde van het gemiddelde van *X* ligt

3. Gemiddeld genomen werkt een Nederlander 30 uur per week. Ga ervanuit dat deze variabele normaal verdeeld is met een standaarddeviatie van 3. Hoe groot is dan ongeveer het deel van de Nederlanders dat tussen de 24 en 36 uur werkt?

A. 5%
B. 32%
C. 68%
D. 95%

4. Rimmer doet een onderzoek naar de gemiddelde tevredenheid van Pedagogiekstudenten met hun tentamencijfer op statistiek. Hij gebruikt daarbij een schaal van 0 tot 100 en gaat ervan uit dat de scores normaal verdeeld zullen zijn. Rimmer steelt een 95% betrouwbaarheidsinterval op voor het gemiddelde uit een random steekproef. Het betrouwbaarheidsinterval loopt van 60 tot 75. Wat betekent dit interval?

A. 95% van de scores in de steekproef liggen tussen de 57 en 63
B. 95% van de scores in de populatie liggen tussen de 57 en 63
C. Er is 95% kans dat dit interval het populatiegemiddelde bevat
D. Er is 95% kans dat dit interval het steekproefgemiddelde bevat

5. Aan 100 Groningse studenten is gevraagd hoeveel biertjes zij de afgelopen week hebben gedronken. De scores zijn rechtsscheef verdeeld met een gemiddelde van 5 en een standaarddeviatie van 3. Hoe veel biertjes moet een student drinken om bij de hoogste 2.5% te zitten?

A. Minimaal 8
B. Minimaal 11
C. Minimaal 14
D. Daar kan op basis van deze gegevens geen uitspraak over worden gedaan

6. De scores op een tentamen zijn normaal verdeeld met gemiddelde 60 en standaarddeviatie 8. Wat is de score die je moet behalen om tot de 5% laagste scores te behoren?

A. Ongeveer 44 of lager
B. Ongeveer 44 of hoger
C. Ongeveer 47 of lager
D. Ongeveer 47 of hoger

7. De tijd om een tentamen af te ronden is normaal verdeeld met een gemiddelde van 50 minuten en een standaarddeviatie van 10 minuten. Wat is ongeveer het percentage studenten dat het tentamen binnen een uur afrondt?

A. 68%
B. 84%
C. 95%
D. 99.7%

8. Uit een onderzoek blijkt dat Nederlanders gemiddeld 1200 euro per jaar uitgeven aan kleding, met een standaarddeviatie van 14.83. Gegeven is dat de *margin of error* (m) 30 is. Hoe groot moet de steekproef minimaal zijn om een 95% betrouwbaarheidsinterval op te kunnen stellen?

A. 5
B. 6
C. 33
D. 34

Antwoorden Hoofdstuk 6

1	D	Een betrouwbaarheidsinterval gebruik je om met een bepaalde (on)zekerheid uitspraken te doen over de populatie; je wilt immers uitspraken doen over de populatie. De steekproef is slechts een middel daartoe.
2	D	Zie uitleg bij vraag 1.
3	D	Neem 1 standaarddeviatie links en rechts van het gemiddelde. Volgens de 68-95-99.7 regel heb je daarmee dus op 95% van alle observaties.
4	C	
5	D	Het is eens *linksscheve verdeling*, waardoor je niet zonder meer dit soort uitspraken mag doen: de 68-95-99.7 regel gaat hier niet op.
6	C	Let op: bij de vorige vraag kon je makkelijk schatten met de vuistregel (ongeveer 2 standaarddeviaties, dus z = 2), maar nu moet je de z-score echt opzoeken. Laagste 5% betekent dat je moet kijken in tabel A bij een linkeroverschrijdingskans van .05. Deze ligt tussen -1.64 en -1.65 in, dus z = -1.645. Het antwoord is dan: $60 - 1.645 * 8 = 46.84$. Dit is afgerond 47.
7	B	Binnen een uur, betekent +1 standaarddeviatie naar rechts. 1 SD aan beide kanten omvat 68%. Tel daar de helft van de resterende 32% bij op, dus: 68 + 16 = 84%.
8	D	$n = (\frac{z*\sigma}{m})^2 = (\frac{1.96*14.83}{5})^2 = (\frac{29.07}{5})^2 = 5.813^2 = 33.80$ dus minimaal 34

Hoofdstuk 7

1. Gegeven zijn twee onafhankelijke variabelen X en Y. Verder is bekend dat het gemiddelde van *X* gelijk is aan 20 en de standaarddeviatie gelijk is aan 10. Variabele *Y* heeft een gemiddelde van 10 en een standaarddeviatie van 5. Wat is de standaarddeviatie van de variabele (*X* – *Y*)?

A. 5
B. 15
C. 75
D. 125

2. Stel we hebben twee onafhankelijke random variabelen X en Y. Welke van onderstaande uitspraken is niet juist?

A. De variantie van het verschil X – Y is gelijk aan het verschil van de varianties
B. De variantie van de som X + Y is gelijk aan de som van de varianties
C. Het gemiddelde van de som X + Y is gelijk aan de som van de gemiddelden
D. Het gemiddelde van het verschil X – Y is gelijk aan het verschil van de gemiddelden

Antwoorden Hoofdstuk 7

1	D	De variantie van de verschil variabele is gelijk aan de som van de varianties van beide variabelen. Dus $var(X - Y) = var(X) + var(Y) = 10^2 + 5^2 = 125.$
2	A	

Niet in het boek; aanvullend materiaal bij Hoofdstuk 2 (collegestof)

1. Twee beoordelaars hebben een groot aantal stimuli op een bepaald kenmerk beoordeeld. De oordelen van deze beoordelaars zijn gegeven als scores op de variabelen X en Y. Om Kendall's tau uit te rekenen tussen X en Y wordt gewerkt met concordante en discordante paren. Wat is een concordant paar?
   A. Een paar stimuli die onderling gelijk zijn.
   B. Een paar beoordelaars die de stimuli gelijke scores geven op een bepaald kenmerk.
   C. Een paar variabelen waarvoor de correlatie tussen de scores positief is.
   D. Een paar stimuli waarvoor de ordening van de X-scores gelijk is aan de ordening van de Y-scores.

2. Spearman's rho geeft aan in hoeverre
   A. De scores op twee variabelen gelijk zijn aan elkaar
   B. Er een lineair verband is tussen de scores op twee variabelen
   C. De absolute waarden van de scores op twee variabelen gelijk zijn
   D. Er overeenstemming is in de rangorde van scores op twee variabelen

3. Wat weet je zeker als Kendall's tau tussen variabelen *X* en *Y* precies 1.0 is?
   A. Iedere proefpersoon heeft op *X* hetzelfde gescoord als op *Y*
   B. Iedere proefpersoon heeft op *X* precies 1 punt hoger gescoord dan op *Y*
   C. De Spearman's correlatie tussen *X* en *Y* is 1.0
   D. De volgorde van de proefpersonen wanneer ze geordend zouden worden op *X* is precies gelijk aan de volgorde van de proefpersonen wanneer ze geordend zouden worden op *Y*.

4. Kan de kappa negatief zijn?
   A. Nee
   B. Ja, maar alleen als de geobserveerde proportie overeenstemming negatief is
   C. Ja, maar alleen als de geobserveerde proportie overeenstemming lager dan 0.50 is
   D. Ja, maar alleen als de geobserveerde proportie overeenstemming lager is dan de verwachte proportie overeenstemming

5. Van 6 doosjes zijn de hoogte (*H*) en lengte (*L*) bepaald. Het blijkt dat in 15 onderlinge vergelijkingen van de doosjes, in 12 gevallen het langste doosje ook het hoogste is. Bij 3 gevallen was het langste doosje juist het laagste. Wat is de Kendall's tau tussen *H* en *L*?
   A. 0.20
   B. 0.25
   C. 0.60
   D. 0.80

Antwoorden bij aanvullend materiaal

1	D	
2	D	
3	D	Kendall's tau zegt enkel iets over de rangorde. Een gelijke ordening betekent niet dat de absolute scores gelijk zijn, of dat ze op één lijn liggen. Het zegt ook niks over het verschil tussen scores.
4	D	De kappa heeft een range van -1 tot 1 en is negatief wanneer de geobserveerde proportie overeenstemming lager is dan de verwachte proportie overeenstemming.
5	C	Kendall's tau = $\frac{\#C-\#D}{aantal\ paren} = \frac{12-3}{15} = 0.60$