

voorbeeldexamenopgaven statistiek wiskunde A havo

FORMULEBLAD

Vuistregels voor de grootte van het verschil van twee groepen

$$2 \times 2 \text{ kruistabel } \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \text{ met } \phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

- als $\phi < -0,4$ of $\phi > 0,4$, dan zeggen we “het verschil is groot”,
- als $-0,4 \leq \phi < -0,2$ of $0,2 < \phi \leq 0,4$, dan zeggen we “het verschil is middelmatig”,
- als $-0,2 \leq \phi \leq 0,2$, dan zeggen we “het verschil is gering”.

Maximaal verschil in cumulatief percentage ($\max V_{cp}$) (met steekproefomvang $n > 100$)

- als $\max V_{cp} > 40$, dan zeggen we “het verschil is groot”,
- als $20 < \max V_{cp} \leq 40$, dan zeggen we “het verschil is middelmatig”,
- als $\max V_{cp} \leq 20$, dan zeggen we “het verschil is gering”.

Effectgrootte $E = \frac{\bar{X}_1 - \bar{X}_2}{\frac{1}{2}(S_1 + S_2)}$, met \bar{X}_1 en \bar{X}_2 de steekproefgemiddelden

($\bar{X}_1 \geq \bar{X}_2$), S_1 en S_2 de steekproefstandaardafwijkingen

- als $E > 0,8$, dan zeggen we “het verschil is groot”,
- als $0,4 < E \leq 0,8$, dan zeggen we “het verschil is middelmatig”,
- als $E \leq 0,4$, dan zeggen we “het verschil is gering”.

Twee boxplots vergelijken

- als de boxen¹⁾ elkaar niet overlappen, dan zeggen we “het verschil is groot”,
- als de boxen elkaar wel overlappen en een mediaan van een boxplot buiten de box van de andere boxplot ligt, dan zeggen we “het verschil is middelmatig”,
- in alle andere gevallen zeggen we “het verschil is gering”.

noot 1 De ‘box’ is het interval vanaf het eerste kwartiel tot en met het derde kwartiel.

Betrouwbaarheidsintervallen

Het 95%-betrouwbaarheidsinterval voor de populatieproportie is

$p \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$, met p de steekproefproportie en n de steekproefomvang.

Het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde is

$\bar{X} \pm 2 \cdot \frac{S}{\sqrt{n}}$, met \bar{X} het steekproefgemiddelde, n de steekproefomvang en S de steekproefstandaardafwijking.

Kijkcijfers

Op 5 juni 2014 keek 9,2% van de Nederlandse bevolking van 6 jaar en ouder op televisie naar het NOS-journaal van 20.00 uur.

We zeggen dan dat de **kijkdichtheid** van dit journaal 9,2 is. Stel dat dit cijfer het resultaat is van een kijkonderzoek onder een aselechte steekproef van 2700 personen van 6 jaar en ouder.

- 4p 1 Laat zien dat het 95%-betrouwbaarheidsinterval voor de kijkdichtheid van dit journaal gelijk is aan $9,2 \pm 1,1$.

Stichting Kijk Onderzoek (SKO) houdt het kijkonderzoek om uitspraken te kunnen doen over hoeveel en welke mensen van de Nederlandse bevolking van 6 jaar en ouder naar de verschillende televisiezenders en programma's kijken.

SKO wil dat het betrouwbaarheidsinterval smaller wordt. Hiervoor moet het aantal deelnemers aan het kijkonderzoek worden vergroot.

- 4p 2 Hoe groot moet de steekproef minstens zijn, zodat de totale breedte van het 95%-betrouwbaarheidsinterval in het kijkonderzoek kleiner dan 1 is? Geef je antwoorden in tientallen nauwkeurig.

In werkelijkheid wordt voor het kijkonderzoek geen aselechte steekproef van personen van 6 jaar en ouder genomen. We bekijken verschillende mogelijkheden om een steekproef te trekken.

Een mogelijke aanpak zou zijn om eerst een aselechte steekproef uit de Nederlandse adressen te trekken. Uit de bewoners op elk gekozen adres wordt vervolgens aselekt één deelnemer aan het kijkonderzoek getrokken.

- 2p 3 Leg uit waarom je hiermee geen aselechte steekproef van personen van 6 jaar en ouder hebt getrokken.

Een andere mogelijkheid is dat SKO een grote aselechte steekproef van personen van 6 jaar en ouder trekt en dan aan deze personen vraagt om zich zelf op te geven voor het kijkonderzoek. Dat levert gemotiveerde deelnemers op.

- 2p 4 Leg uit waarom deze aanpak voor SKO niet verstandig is.

Voor de werkelijke trekking van deelnemers aan het kijkonderzoek worden de deelnemers geselecteerd door eerst een steekproef van Nederlandse gemeenten te trekken. Dit wordt gedaan door de gemeenten op een lijst te zetten, gesorteerd op volgorde van grootte. Met vaste stapgrootte wordt door de lijst gegaan en de namen van de getrokken gemeenten worden genoteerd. Uit deze gemeenten zullen vervolgens de deelnemers worden getrokken.

- 2p 5 Geef aan hoe deze eerste stap de kwaliteit van de steekproef van deelnemers bevordert.

Verschillen

Op de Universiteit van Leiden is een onderzoek gedaan onder een groep studenten. Hierbij werden onder andere de lichaamslengte en de armlengte van de studenten gemeten. Ook werden het geslacht, de kleur van de ogen en de voorkeurshand genoteerd. In totaal ging het om 34 mannelijke studenten en 32 vrouwelijke studenten. Enkele resultaten zie je in tabel 1. Op basis van resultaten van dit onderzoek wil men uitspraken doen over alle mannelijke en vrouwelijke studenten in Leiden.

tabel 1

		voorkeurshand	
		links	rechts
geslacht	man	5	29
	vrouw	2	30

- 4p **6** Noem de variabelen uit tabel 1 en geef van elk van de variabelen aan of deze nominaal of ordinaal is. Licht je antwoord toe.
- 3p **7** Bepaal met behulp van het formuleblad of het verschil in voorkeurshand tussen de mannelijke en de vrouwelijke studenten in het onderzoek groot, middelmatig of gering is.

In tabel 2 staan het gemiddelde en de standaardafwijking van de lichaamslengte van de studenten.

tabel 2

	aantal	lichaamslengte (cm)	
		gemiddelde	standaardafwijking
man	34	183,8	5,8
vrouw	32	170,8	8,1

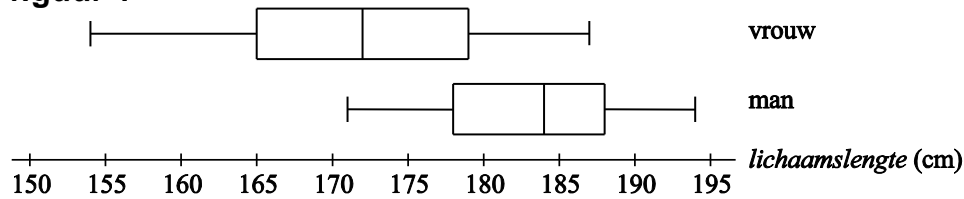
- Op basis van tabel 2 kunnen voor de lichaamslengte van mannen en vrouwen 95%-betrouwbaarheidsintervallen worden opgesteld.
- 4p **8** Onderzoek welk van deze twee intervallen het smalst is. Welke informatie geeft zo'n 95%-betrouwbaarheidsinterval in deze context?

Iemand die iets wil zeggen over de grootte van het verschil in lichaamslengte tussen de mannen en de vrouwen in het onderzoek, zou gebruik kunnen maken van de effectgrootte. Bij de berekening van de effectgrootte zijn de gemiddeldes en standaardafwijkingen van beide groepen van belang. Ga ervan uit dat het verschil van de gemiddeldes gelijk blijft.

- 3p **9** Leg uit op welke wijze grotere standaardafwijkingen de grootte van het verschil (groot, middelmatig, gering) kunnen beïnvloeden.

De resultaten van de lichaamslengten werden ook in boxplots verwerkt. Zie figuur 1.

figuur 1



Je kunt met behulp van de gegevens over de lichaamslengten in tabel 2 onderzoeken hoe groot het verschil in lichaamslengte in het onderzoek is. Dit kun je ook doen met behulp van de gegevens in figuur 1.

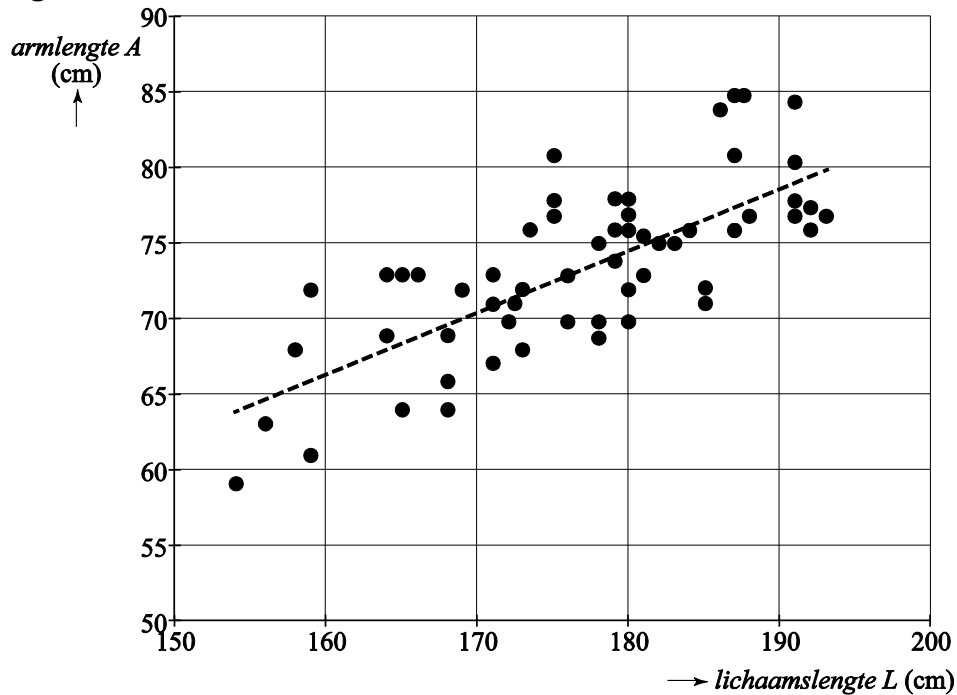
- 6p **10** Toon met behulp van het formuleblad aan dat je met deze twee gegevensbronnen een verschillende conclusie trekt over de grootte van het verschil in lichaamslengte tussen de mannelijke en de vrouwelijke studenten.

- 3p **11** Beargumenteer welk van beide conclusies het best te verdedigen is.

Om te onderzoeken of er een verband bestaat tussen de lichaamslengte en de armlengte van de studenten zijn de resultaten van het onderzoek in een spreidingsdiagram gezet. Zie figuur 2.

Er blijkt in dit onderzoek bij benadering een lineair verband te zijn tussen de lichaamslengte L in cm en de armlengte A in cm. De getekende trendlijn geeft dit verband weer.

figuur 2



- 4p **12** Stel een formule op van de trendlijn.

De punten in het spreidingsdiagram liggen niet precies op de trendlijn. Ga ervan uit dat de armlengte normaal verdeeld is bij alle mogelijke lichaamslengten tussen 150 cm en 200 cm en dat bij elke lichaamslengte de standaardafwijking van de bijbehorende armlengte gelijk is aan 3,0 cm. Als we de resultaten van dit onderzoek gebruiken om op basis van de lichaamslengte de armlengte van studenten te voorspellen, kunnen we hierbij figuur 2 gebruiken.

- 4p **13** Teken in figuur 2 het gebied waar de middelste 95% van de armlengten ligt voor elk van de lichaamslengten tussen 150 en 200 cm.

Hommels

Er is onderzoek gedaan naar het gedrag van hommels. De onderzoekers vroegen zich hierbij af hoe efficiënt hommels van bloem naar bloem vliegen om nectar te vinden.



Tijdens het onderzoek werd steeds een hommelm in een afgesloten ruimte geplaatst. In deze ruimte bevonden zich zes kunstmatige bloemen met daarin een nectarachtige suikeroplossing. Telkens als de hommelm vanuit het hommelnest op zoek ging naar voedsel, werd na terugkeer op het nest genoteerd welke route hij die vlucht volgde, welke afstand werd afgelegd en hoelang de vlucht duurde.

De onderzoekers deden het onderzoek met 8 verschillende hommels. Van elke hommelm werden de gegevens genoteerd van 8 series van elk 10 vluchten. Dat leverde in totaal 640 onderzoeksresultaten op. Niet alle onderzoeksresultaten waren bruikbaar. Als de hommelm niet alle zes bloemen had bezocht, werd de vlucht afgekeurd. In tabel 1 zie je per hommelm het aantal afgekeurde vluchten.

tabel 1 Aantal afgekeurde vluchten

	hommelm								totaal
	1	2	3	4	5	6	7	8	
serie A: vlucht 1 t/m 10	4	1	3	10	2	4	1	4	29
serie B: vlucht 11 t/m 20	2	0	0	0	3	2	2	1	10
serie C: vlucht 21 t/m 30	4	2	0	0	1	1	0	1	9
serie D: vlucht 31 t/m 40	0	1	1	0	1	0	0	2	5
serie E: vlucht 41 t/m 50	3	0	2	0	2	0	1	0	8
serie F: vlucht 51 t/m 60	0	1	0	0	3	0	0	0	4
serie G: vlucht 61 t/m 70	1	0	0	0	0	0	0	0	1
serie H: vlucht 71 t/m 80	0	0	0	0	0	0	0	0	0
totaal	14	5	6	10	12	7	4	8	66

In tabel 1 zie je bij hommelm 1 bijvoorbeeld dat 4 van de eerste 10 vluchten zijn afgekeurd. In de laatste serie van 10 vluchten werd van deze hommelm geen enkele vlucht afgekeurd.

De onderzoekers concludeerden op grond van de tabel: “De hommels hadden in het begin wat moeite om alle bloemen te vinden. Het percentage afgekeurde vluchten in serie A is ... keer zo hoog als het percentage afgekeurde vluchten in het gehele onderzoek.”

4p 14 Bereken welk getal er op de puntjes moet staan.

De onderzoekers vergeleken de afgelegde afstand van de routes van de goedgekeurde vluchten in de verschillende series. Zie tabel 2.

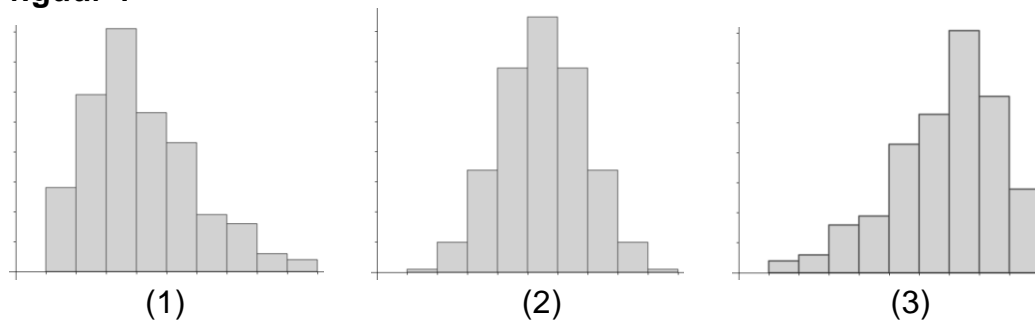
tabel 2

		serie A: vlucht 1 t/m 10	serie B: vlucht 11 t/m 20	serie H: vlucht 71 t/m 80
aantal goedgekeurde vluchten		51	70	80
aantal keer kortste route		0	4	35
afgelegde afstand per goedgekeurde vlucht (cm)	gemiddelde	6541	5473	3840
	standaardafwijking	1354	1187	512
	mediaan	5419	4892	3329

De onderzoekers keken naar de afgelegde afstand per goedgekeurde vlucht en stelden: "Vaker vliegen maakt verschil. In de laatste serie is de gemiddelde afgelegde afstand per goedgekeurde vlucht een stuk kleiner dan in de eerste serie."

- 3p 15 Bepaal met behulp van het formuleblad of het verschil groot, middelmatig of gering is.

figuur 1

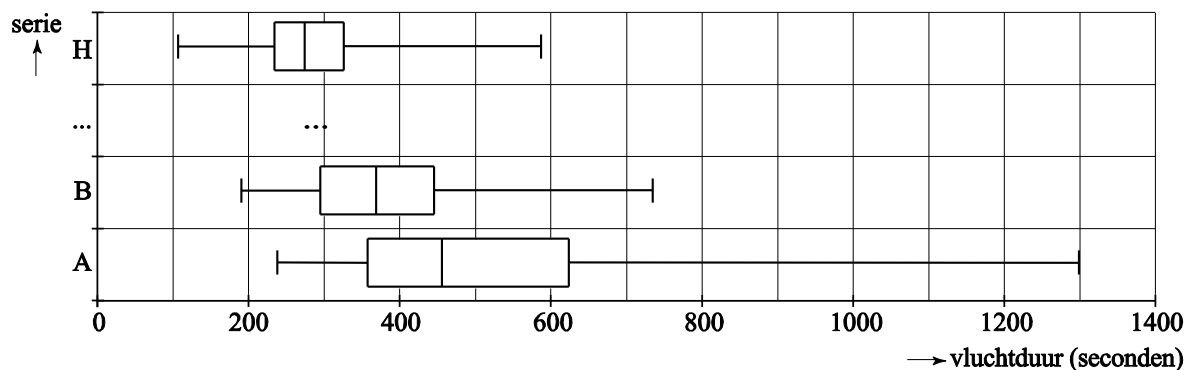


In figuur 1 zie je drie schetsen van een verdeling.

- 3p 16 Leg uit welk van deze schetsen het best past bij de de afgelegde afstand per goedgekeurde vlucht van serie A.

Ook de vluchtduur (het aantal seconden dat een vlucht duurt) van alle goedgekeurde vluchten is gemeten. In de figuur zie je voor de series A, B en H de boxplot van de vluchtduur in seconden.

figuur 2



Een vergelijking van de boxplots laat zien dat de vluchtduur afneemt. De onderzoekers vragen zich af of er al verschil is in vluchtduur tussen de vluchten in serie A en die in serie B.

- 3p 17 Bepaal met behulp van het formuleblad hoe groot het verschil in vluchtduur is tussen serie A en serie B.

Blijkbaar weten de hommels de weg naar de zes bloemen steeds beter te vinden. Veel hommels hebben vroeg of laat de kortste route van 2462 cm gevonden.

De hommel die in serie H de snelste tijd neerzette, nam deze kortste route. Deze hommel had tijdens zijn vlucht gemiddeld 5 seconden per bloem nodig om de suikeroplossing eruit te halen. De overige tijd gebruikte hij om zijn route langs de zes bloemen te vliegen.

- 3p 18 Gebruik figuur 2 om de gemiddelde vliegsnelheid van deze hommel tijdens zijn snelste vlucht te berekenen. Geef je antwoord in cm per seconde.

Met de informatie uit het onderzoek kun je op verschillende manieren de conclusie trekken dat hommels kunnen leren.

- 3p 19 Geef drie argumenten om deze conclusie te ondersteunen en vermeld bij elk argument uit welke tabel of figuur de informatie komt.

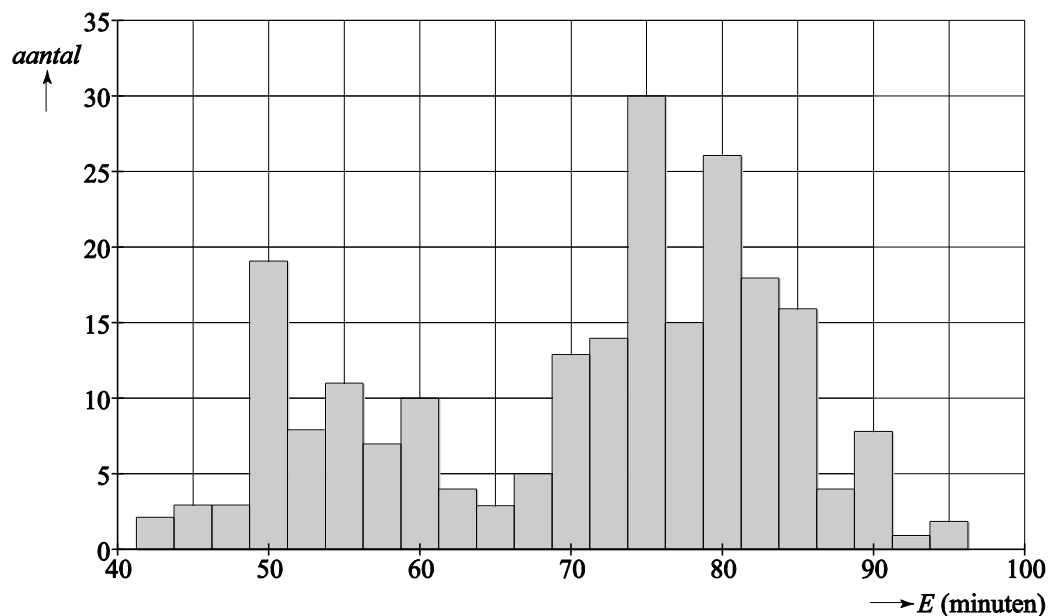
Old Faithful

Een geiser is een heetwaterbron, die met regelmatige tussenpozen heet water en stoom de lucht in spuit. De Old Faithful geiser in het Yellowstone park in de Verenigde Staten is één van de bekendste geisers ter wereld.

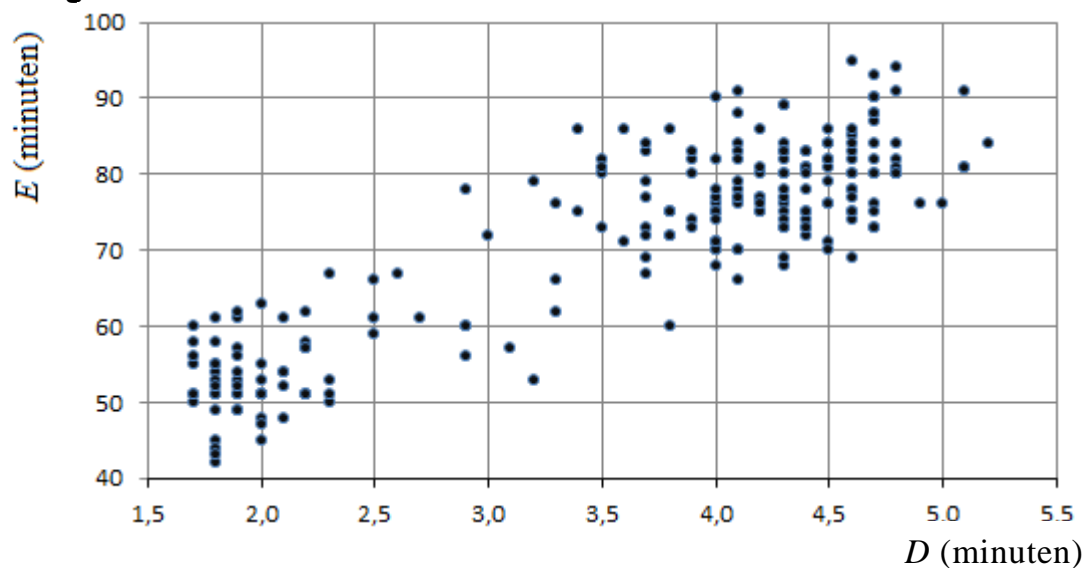


Van 222 opeenvolgende uitbarstingen van de Old Faithful zijn gegevens verzameld over de tijdsduur van iedere uitbarsting D en de tussentijd tot de eerstvolgende uitbarsting E . Hierbij zijn D en E in minuten. Deze gegevens zijn verwerkt in de figuren 1 en 2.

figuur 1



figuur 2



De tussentijden tot de eerstvolgende uitbarsting zijn te verdelen in kort en lang.

4p **20** Geef op basis van figuur 1 een schatting van de verhouding tussen het aantal korte en lange tussentijden. Licht je werkwijze toe.

3p **21** Hoe lang is de tijdsduur van de beide uitbarstingen die horen bij de twee kleinste tussentijden tot de eerstvolgende uitbarsting? Licht je werkwijze toe.

In een toeristische folder over de Old Faithful staat onder andere de volgende informatie:

- 1 De tijdsduur van een uitbarsting varieert tussen de 1,5 en 5,5 minuten.
- 2 Het kortste tijdsinterval tussen twee uitbarstingen was ongeveer 42 minuten.
- 3 De hoogte van het water is bij elke uitbarsting tussen 32 en 56 meter.
- 4 Hoe langer een uitbarsting duurt, hoe langer het duurt tot de volgende uitbarsting.

Hieronder staat een tabel waarin je kunt invullen of je bovenstaande informatie uit de figuren 1 en 2 kan aflezen.

tabel

informatie	af te lezen uit figuur 1 of figuur 2 of uit beide figuren?
1	
2	
3	
4	

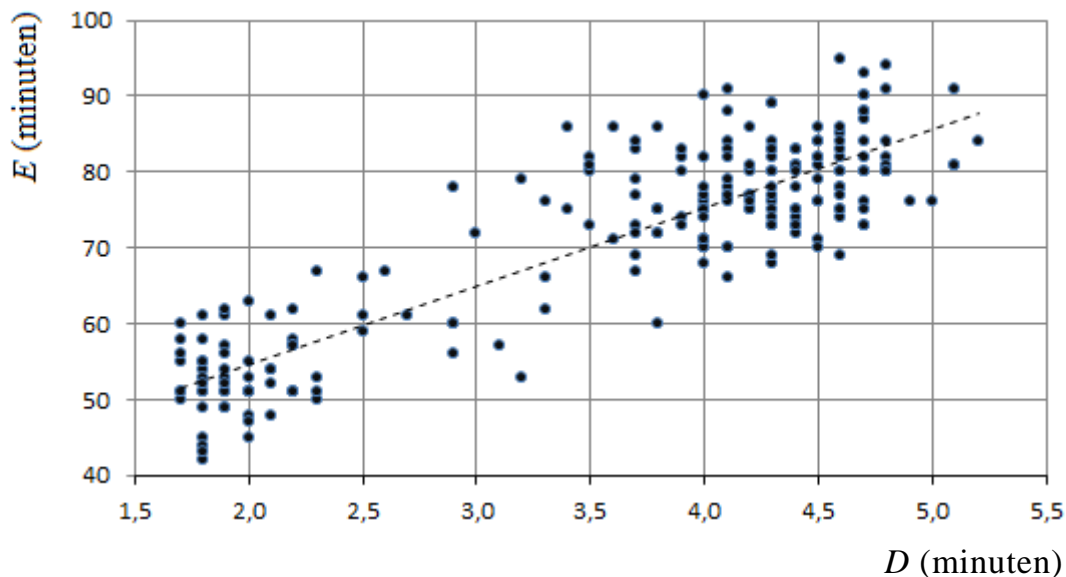
3p **22** Vul de tabel in.

Je kunt met behulp van figuur 2 schatten dat de gemiddelde tussentijd tot de eerstvolgende uitbarsting ongeveer 70 minuten is.

4p **23** Gebruik figuur 2 om een schatting te geven van het gemiddeld aantal uitbarstingen van de Old Faithful per dag. Licht je werkwijze toe.

Door de regelmaat van de uitbarstingen van de Old Faithful zijn de uitbarstingen vrij goed voorspelbaar. In figuur 3 staat nogmaals figuur 2 met daarin de trendlijn.

figuur 3



De formule van de trendlijn is

$$E = 10D + 35, \text{ met } D \text{ en } E \text{ in minuten.}$$

3p **24** Toon dit aan.

Door de regelmaat van de uitbarstingen van de Old Faithful kan na een uitbarsting met behulp van de formule het tijdstip van de volgende uitbarsting worden voorspeld aan de hand van het tijdstip en de duur van de laatste uitbarsting. Uit de praktijk is gebleken dat de voorspelling van 99% van deze tijdstippen niet meer dan 15 minuten van het werkelijke tijdstip verschilt.

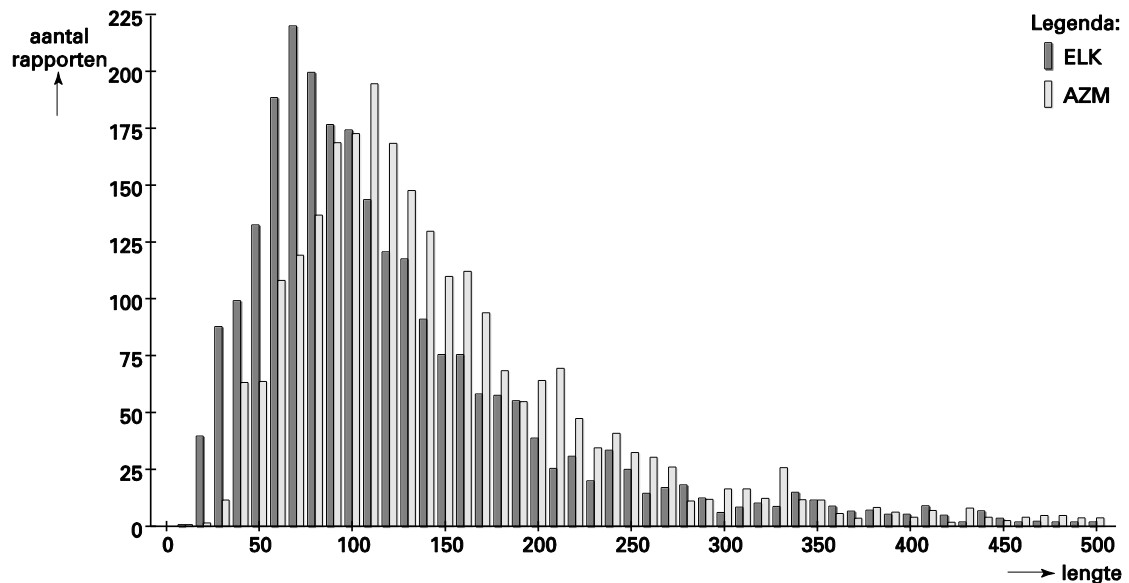
Een toerist komt om 11.07 uur aan bij de ingang van Yellowstone park. Hij hoort van een parkwachter dat hij net een uitbarsting van 4,3 minuten heeft gemist. De toerist gaat wandelen in het park, maar hij wil de eerstvolgende uitbarsting niet missen.

3p **25** Bereken met behulp van de trendlijn op welk tijdstip de toerist bij de Old Faithful moet zijn om de volgende uitbarsting met 99% zekerheid te zien.

Teksten vergelijken (n.a.v. vwo A1 2004-II)

In ziekenhuizen worden vaak medische rapporten geschreven. Bij een onderzoek naar de inhoud van dergelijke rapporten zijn 2500 rapporten van het Elkerliek Ziekenhuis (ELK) te Deurne vergeleken met 2500 rapporten van het Academisch Ziekenhuis Maastricht (AZM). Van elk rapport is de lengte bepaald, dat is het aantal woorden dat het rapport bevat. In figuur 1 zijn de gegevens weergegeven in een gecombineerd staafdiagram met klassenbreedte 10.

figuur 1



Voor de lengte van de rapporten van het ene ziekenhuis geldt:

I: Het 1e kwartiel is 68, de mediaan is 100 en het 3e kwartiel is 149.

Voor de lengte van de rapporten van het andere ziekenhuis geldt:

II: Het 1e kwartiel is 92, de mediaan is 127 en het 3e kwartiel is 184.

- 3p **26** Welke van deze series gegevens, I of II, hoort bij het ELK? Licht je antwoord toe.

Uit het onderzoek bleek dat de mediaan en het gemiddelde van de lengte van de rapporten van het AZM niet even groot zijn.

- 3p **27** Geef met een redenering, dus zonder een berekening, aan of de mediaan groter of kleiner is dan het gemiddelde.

De rapporten van beide ziekenhuizen bevatten samen 996 734 woorden. Toch waren er in totaal slechts ongeveer 20 000 verschillende woorden. Dit komt omdat er woorden zijn die heel vaak gebruikt worden. In de tabel zie je de 10 woorden die het meest frequent in de rapporten werden gebruikt.

tabel Meest gebruikte woorden in medische teksten

woord	een	de	van	met	en	het	in	is	ik	geen
rangnummer	1	2	3	4	5	6	7	8	9	10
frequentie	40 361	36 485	34 231	27 667	26 869	22 965	22 082	13 681	11 416	11 363
cumulatief percentage	4,0	7,7	11,1	13,9	16,6	18,9	21,1	22,5	23,7	24,8

Je ziet dat in de tabel de woorden op rangnummer, in volgorde van hun frequentie, zijn genoemd. Zo kun je bijvoorbeeld aflezen dat het woord 'met' in totaal 27 677 keer is geteld en dat dit woord rangnummer 4 heeft.

Op grond van het onderzoek kun je berekenen hoeveel keer per 10 000 woorden het woordje 'geen' in een medisch rapport gebruikt wordt. Ook kun je het bijbehorende 95%-betrouwbaarheidsinterval berekenen.

- 4p **28** Bereken dit aantal keer per 10 000 woorden en bepaal het bijbehorende 95%-betrouwbaarheidsinterval.

Voor algemene teksten is er een formule die het verband aangeeft tussen het rangnummer r van een woord en de bijbehorende frequentie f . Deze formule staat bekend als de wet van Zipf.

In het geval van een algemene tekst van 996 734 woorden is de formule:

$$f = \frac{88000}{r} \quad (\text{wet van Zipf})$$

Medische rapporten bevatten uiteraard medische teksten. Men wil onderzoeken in hoeverre er verschil is tussen medische teksten en algemene teksten.

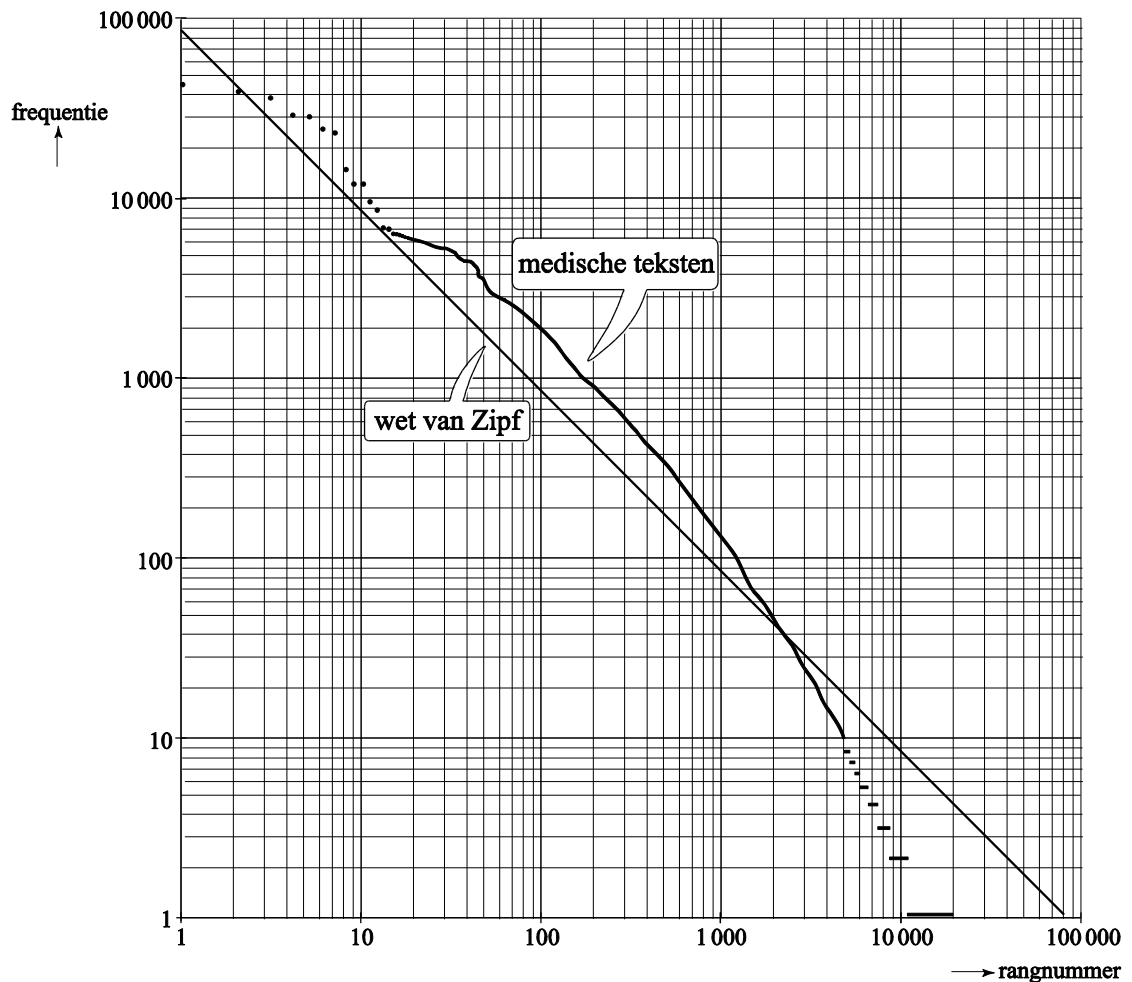
Daartoe worden voor de medische teksten de cumulatieve percentages van de 10 meest gebruikte woorden gebruikt. Zie de tabel. Met behulp van de wet van Zipf kan men ook cumulatieve percentages van de meest gebruikte woorden in algemene teksten bepalen.

Op grond van de cumulatieve frequenties van beide teksten en een vuistregel op het formuleblad kan dan worden bepaald in hoeverre er verschil is tussen de onderzochte medische teksten en algemene teksten. Op den duur lopen deze cumulatieve percentages behoorlijk gelijk, maar in het begin wijken ze behoorlijk van elkaar af. Hierdoor is het voldoende naar de eerste 3 rangnummers te kijken.

- 5p **29** Onderzoek op bovenstaande wijze of het verschil tussen de onderzochte medische teksten en algemene teksten gering, middelmatig of groot is.

Men wil de resultaten van het onderzoek nog wat preciezer bekijken. Daarom zijn in figuur 2 van alle gebruikte woorden in de medische teksten de frequenties uitgezet tegen de rangnummers. Op beide assen is gekozen voor een logaritmische schaalverdeling. De woorden uit de tabel vind je terug als de bovenste 10 punten. Ook is de grafiek van de wet van Zipf getekend.

figuur 2



Je ziet dat, behalve van de woorden 'een' en 'de', de frequenties van de 2000 meest gebruikte woorden in de medische teksten hoger zijn dan verwacht mag worden volgens de wet van Zipf.

- 4p **30** Bereken hoeveel procent het woord met rangnummer 100 in medische teksten vaker voorkomt dan volgens de wet van Zipf verwacht mag worden.

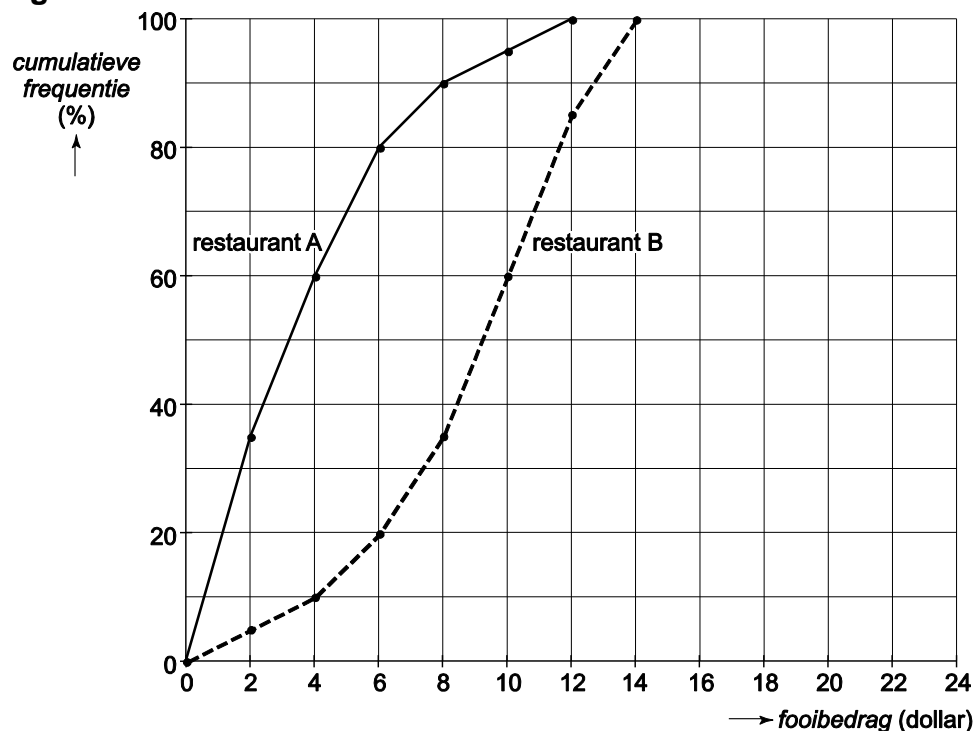
Foien (n.a.v. havo A1,2 2006-II)

In de Verenigde Staten is het gebruikelijk dat je in een restaurant een flinke fooi geeft aan degene die je bedient. Het basisloon is er zeer laag en daardoor is het bedienend personeel veel meer afhankelijk van foien dan in Nederland.

In New York bestudeerde een onderzoeker welke foien er gegeven werden bij 500 rekeningen in restaurant A en 500 rekeningen in restaurant B.

In de figuur zijn de cumulatieve relatieve frequentiepolygonen van de foien van de restaurants A en B getekend.

figuur



2p 31 In welk restaurant werd in totaal meer fooi gegeven? Licht je antwoord toe.

3p 32 In welk restaurant werden er relatief meer foien tussen de 6 en de 8 dollar gegeven? Licht je antwoord toe.

Ruim driekwart van de foien in restaurant B is hoger dan de 75% laagste foien van restaurant A.

3p 33 Leg uit hoe je dit in de figuur terug kunt vinden.

Op grond van de figuur kun je met behulp van het formuleblad met twee verschillende vuistregels een uitspraak doen over het verschil tussen de fooien in restaurant A en die in restaurant B.

- 6p **34** Beschrijf hoe je met twee verschillende vuistregels dit verschil kunt bepalen en geef bij beide de conclusie: is het verschil groot, middelmatig of gering?