

Statistiek



1 Verschillen

Inhoudsopgave

1.0	Statistiek	4
1.1	Verschillen tussen groepen	6
1.2	Data in beeld	12
1.3	Computerpracticum	22
1.4	Onderzoek 1 / Spiekbriefje	24
1.5	Samenvatting paragraaf 1.1 t/m 1.4	25
1.6	Verschillen meten	27
1.7	Gemiddelde en standaarddeviatie	32
1.8	Effectgrootte	39
1.9	Onderzoek 2	44
1.10	Samenvatting paragraaf 1.6 t/m 1.9	45

Bij dit hoofdstuk hoort een *digimap*. Daarin staan opgaven waarbij er iets met ict valt te beleven. Ze hebben hetzelfde nummer als de bijbehorende opgave in deze tekst. Je herkent ze aan het icoontje . Via de digimap kom je bij VU-Statistiek of op een website terecht.

De opgaven in de digimap kunnen op drie manieren worden gebruikt:

- niet,
- aanvullend op de corresponderende opgave in deze tekst,
- vervangend voor de corresponderende opgave in deze tekst.

Colofon

© 2010 cTWO

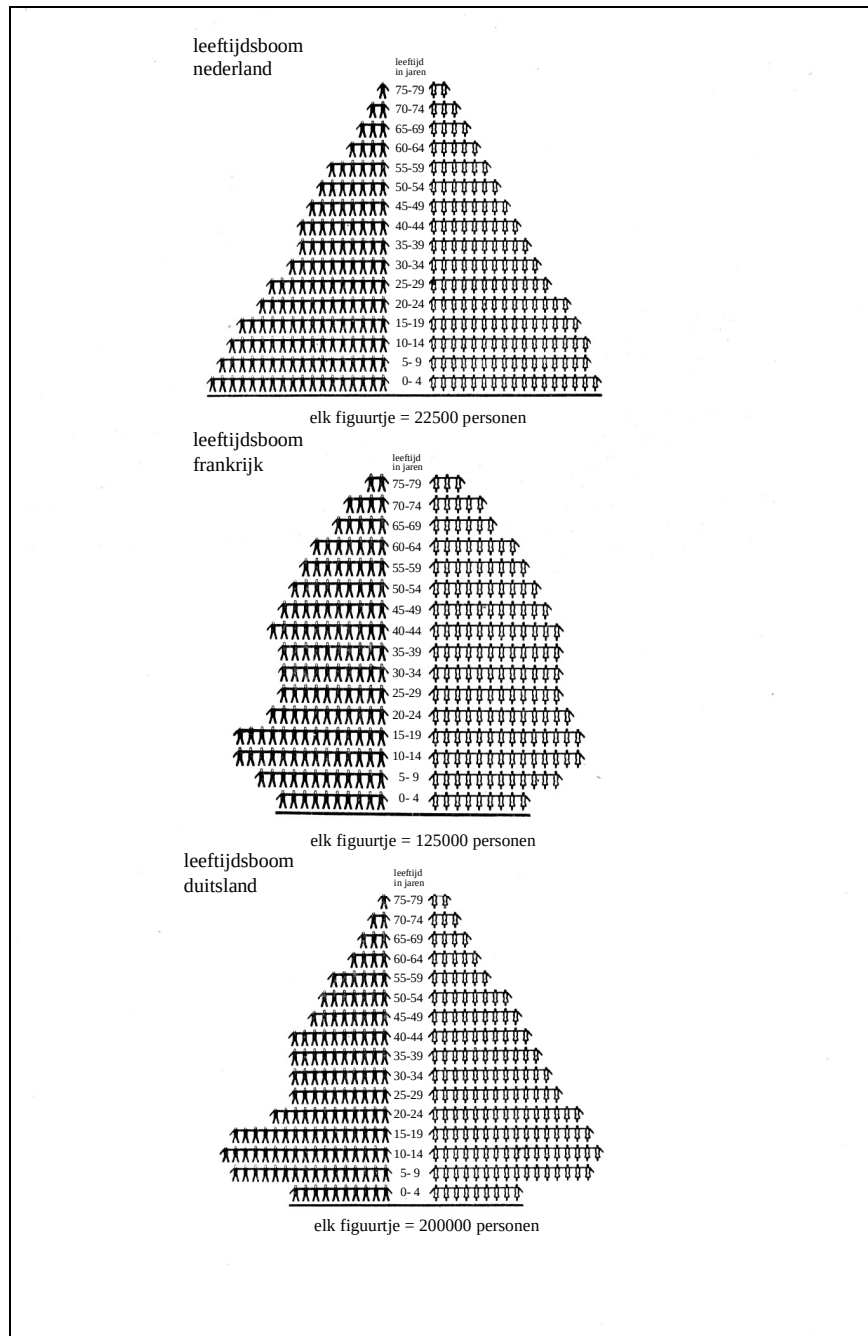
Experimentele uitgave Kansrekening en Statistiek, vwo, wiskunde A en C

versie 2 november 2011

auteurs Leon van den Broek, Maris van Haandel, Bert Nijdam

met medewerking van Simon Biesheuvel, Carel van de Giessen, Sieb Kemme, Peter Kop,
Piet Versnel, Peter van Wijk





Vergelijk de bevolkingspiramides uit 1932 van Nederland, Frankrijk en Duitsland.

Nederland is een braaf groeiend landje.

Uit de piramides is een dramatisch stukje geschiedenis af te lezen. Kijk maar eens naar de leeftijdsklassen tussen 20 en 40 jaar van Frankrijk en Duitsland.

Bron: B.G.Escher, *De methodes der grafische voorstelling*, 1934

1.0 Statistiek

1 We bekijken drie beweringen.

Onze zomers worden steeds natter.

Anneke beweert dat de zomers in Nederland steeds natter worden. Want vorig jaar nog was haar vakantie in eigen land helemaal verregend.

a Wat vind je van Annekes argument?

Meisjes moeten zich niet laten vaccineren tegen baarmoederhalskanker.

Anneke laat zich niet inenten tegen baarmoederhalskanker. Haar vriendinnen doen dat ook niet. Want zij heeft gelezen dat in de VS meisjes ernstig ziek zijn geworden van de inenting.

b Wat vind je van Annekes argument?

Het fileprobleem is veel erger geworden.

Anneke beweert dat het fileprobleem de laatste jaren veel erger is geworden. Want gisteren en eergisteren heeft ze twee uur in de file gestaan tussen Amsterdam en 's-Hertogenbosch.

c Wat vind je van Annekes argument?



wat is statistiek?

Om verantwoorde uitspraken te kunnen doen in onzekere situaties, is statistiek nodig.

- Je begint met een vraag of een bewering.
- Dan moet je keuzes maken. Bijvoorbeeld wat het fileprobleem betreft.
 - Hoever ga je terug in de tijd? Onderscheid je verschillende seizoenen?
 - Kijk je alleen naar Nederland of betrek je ook andere landen in je onderzoek? Onderscheid je verschillende delen van Nederland?
 - Let je op het aantal files per dag, of hoe lang ze zijn, of ...
 - Let je op oorzaken van files? Wegwerkzaamheden, weer, vakanties, ...
 - Als je gemiddeldes neemt, doe je dat dan per dag, per week, per maand?
 - Wat is de invloed van economische factoren?
- Vervolgens ga je gegevens verzamelen.
- Dat levert een gigantische verzameling getallen op waar je in het begin geen grip op hebt. Daartoe ga je ze ordenen en in beeld brengen.
- Dan moet je de gegevens analyseren: je berekent bijvoorbeeld gemiddeldes en de spreiding.
- Tot slot worden er conclusies getrokken; je beantwoordt de vraag.

Statistiek is ontstaan in Duitsland in de achttiende eeuw. In Nederland werden voor het eerst in 1802 te Leiden statistiekcolleges gegeven, door de hoogleraar Adriaan Kluit. Het vak heette toen nog *statenkunde* en bestond uit kennis die nodig was voor een goed bestuur van een land, zoals over de bevolking, het leger, het klimaat, de handel, enz.

- 2 Iedereen krijgt wel eens te maken met files.
- a Noem enkele beroepsgroepen die moeten weten hoe het fileprobleem zich ontwikkelt.
 - b Waarom is statistisch onderzoek daarvoor onmisbaar?
 - c Wie doen in Nederland statistisch onderzoek?

Uit Wikipedia (bewerkt)

Statistiek is de wetenschap, de methodiek en de techniek van het verzamelen, bewerken, interpreteren en presenteren van gegevens. Het is een onderdeel van de wiskunde.

Statistici trachten informatie over een populatie te krijgen uit de waarneming van een (meestal) beperkt deel van die populatie, de zogenaamde steekproef. De zo verkregen informatie is uiteraard onvolledig en daardoor onnauwkeurig. Een goede beheersing van deze onnauwkeurigheid is dan ook een essentieel onderdeel van de statistiek. De uitkomsten kunnen voor allerlei aspecten van de wetenschap, de politiek, de economie, de psychologie en sociologie, de media en de samenleving van belang zijn.

Het woord statistiek is afkomstig van de moderne Latijnse benaming *statisticum collegium* (les over staatszaken), waar het Italiaanse woord *statista* van is afgeleid, wat "staatsman" of "politicus" betekende (vergelijk ons woord status) en het Duitse *Statistik*, wat oorspronkelijk analyse van staatsgegevens betekende.

Begrepen?

De politiechef van Amsterdam zegt dat de criminaliteit hoog is onder Marokkanen.

Waarom is statistisch onderzoek hierbij belangrijk?

Hoe zou jij dit statistisch onderzoeken?

1.1 Verschillen tussen groepen

datamatrix

Op de meeste scholen kies je aan het eind van klas 3 voor een profiel met een wiskundevak. In deze paragraaf ga je je bezighouden met de statistische vraag: “In welk opzicht verschillen de leerlingen die wiskunde A of C hebben gekozen van leerlingen die hebben wiskunde B hebben gekozen?”

- 3 In onderstaande tabel bekijken we vijf kenmerken van leerlingen. De tabel staat ook op het werkblad.
Elk van de kenmerken heeft min of meer invloed op de keuze wiskunde A/C of wiskunde B.

	Verschillen tussen leerlingen met wis A/C en wis B				
Kenmerk	geen	gering	matig	tamelijk	groot
Eindcijfer wiskunde in klas 3					
Gem. alle eindcijfers in klas 3					
Leeftijd					
Aantal vreemde talen in pakket					
Belangstelling voor kunst					

- a Kruis op het werkblad per kenmerk aan hoe sterk het verschil is dat je verwacht.
b Bedenk nog twee kenmerken waarop je een duidelijk verschil vermoedt.



Als je wilt onderzoeken of de verschillen tussen de A/C- en B-kiezers inderdaad zijn zoals je vermoedt, dan zul je gegevens moeten verzamelen. Bijvoorbeeld met behulp van de vragenlijst op het werkblad.



- 4 Vul voor jezelf de vragenlijst op het werkblad in.

Op het Amalia College is de vragenlijst voorgelegd aan de 154 vwo4-leerlingen. De resultaten zijn verzameld in een tabel: een zogenaamde **datamatrix**. Hieronder staat een deel van de datamatrix.

<i>leerling</i>	<i>geslacht</i>	<i>gebjaar</i>	<i>gebmd</i>	<i>gewicht</i>	<i>lengte</i>	<i>kunstbel</i>	<i>talen</i>	<i>cijfwis</i>	<i>cijfgem</i>	<i>huiswerk</i>	<i>wisgroep</i>	<i>profiel</i>	<i>leuk</i>	<i>tedoen</i>	<i>past</i>	<i>advies</i>
16	v	91	12	52	170	4	2	8	7,3	10	B	NG	3	1	2	4
17	v	92	1	59	165	3	3	8	7,0	15	A/C	EM	1	2	4	3
18	m	92	3	69	186	2	2	7	7,5	10	B	NG	4	1	4	4
19	m	91	10	75	190	2	2	6	6,5	7	A/C	EM	2	1	4	3
20	m	91	12	80	178	3	2	8	7,0	7	A/C	CM	2	1	4	3
21	v	92	6	61	177	3	2	8	7,4	6	B	NG	4	4	1	4
22	m	92	6	80	189	1	1	7	6,8	1	A/C	EM	3	1	2	4

Boven de zeventien kolommen staan de **variabelen**, in dezelfde volgorde als in de vragenlijst. Variabelen zijn kenmerken of eigenschappen van (in dit geval) een leerling.

Per regel staan van een leerling alle antwoorden, de **scores**.

De eerste variabele is niet de naam van de leerling, maar een nummer dat aan de leerling is gegeven. Die nummers lopen in ons geval van 1 tot en met 154.

- 5 a Hoe oud in jaren en maanden was leerling 18 op 1 januari 2008?
 b Heeft leerling 18 een slank postuur?
 c Vond leerling 18 het gegeven advies over het profiel belangrijk?
 d Welke leerling heeft het meeste talent, nr. 17 of nr. 22?
- 6 Uit de datamatrix kun je van elke leerling de scores terugvinden. Je hebt nu de gegevens, maar je weet nog niet in hoeverre de wiskunde A/C-leerlingen verschillen van de wiskunde B-leerlingen.
 a Welke leerlingen uit het getoonde stukje van de datamatrix hebben wiskunde B?
 b Heeft die ‘groep’ een hoger cijfer voor wiskunde gehaald in de derde klas dan de wiskunde A/C-groep?
 c Op welke twee variabelen verwacht je een duidelijk verschil tussen wiskunde A/C-leerlingen en wiskunde B-leerlingen? Ga na of die verschillen er ook zijn.

kruistabellen

In opgave 3 heb je aangegeven op welke kenmerken de groep leerlingen met wiskunde B vermoedelijk verschilt van de groep met wiskunde A/C. Nu ga je kijken naar de variabele *geslacht*. Hieronder zie je nog eens de datamatrix voor de zeven leerlingen.

<i>leerling</i>	<i>geslacht</i>	<i>gebjaar</i>	<i>gebmd</i>	<i>gewicht</i>	<i>lengte</i>	<i>kunstbel</i>	<i>talen</i>	<i>cijfwis</i>	<i>cijfgem</i>	<i>huiswerk</i>	<i>wisgroep</i>	<i>profiel</i>	<i>leuk</i>	<i>tedoen</i>	<i>past</i>	<i>advies</i>
16	v	91	12	52	170	4	2	8	7,3	10	B	NG	3	1	2	4
17	v	92	1	59	165	3	3	8	7,0	15	A/C	EM	1	2	4	3
18	m	92	3	69	186	2	2	7	7,5	10	B	NG	4	1	4	4
19	m	91	10	75	190	2	2	6	6,5	7	A/C	EM	2	1	4	3
20	m	91	12	80	178	3	2	8	7,0	7	A/C	CM	2	1	4	3
21	v	92	6	61	177	3	2	8	7,4	6	B	NG	4	4	1	4
22	m	92	6	80	189	1	1	7	6,8	1	A/C	EM	3	1	2	4

- 7a Wat is de verhouding A/C : B onder de meisjes?
 b Wat is de verhouding jongens : meisjes in de A/C-groep?

Met een **kruistabel** waarin je de variabele *geslacht* combineert (“kruist”) met de variabele *wisgroep* krijg je een goed overzicht hoe de verdeling meisjes-jongens samenhangt met de keuze A/C-B. Voor de leerlingen uit de datamatrix krijg je de volgende kruistabel.

<i>geslacht</i>	<i>wisgroep</i>		totaal
	A/C	B	
man			4
vrouw			3
totaal	4	3	7

De **frequenties** (aantallen) zijn nog niet allemaal ingevuld.

- c Noteer op het werkblad de ontbrekende frequenties in de kruistabel.

-  **8** Met de gegevens van alle 154 vwo4-leerlingen van het Amalia College krijg je de volgende kruistabel:

<i>geslacht</i>	<i>wisgroep</i>		totaal
	A/C	B	
man	13	56	69
vrouw	30	55	
totaal			154

- a** Vul op het werkblad de ontbrekende aantallen in.
b Vind je dat de A/C- en de B-keuze veel verschillen in de verdeling op de variabele *geslacht*?
Waarom wel/niet?

Voor een goede vergelijking kun je beter nagaan hoeveel jongens en meisjes er *relatief* zijn in beide groepen, door de aantallen te herleiden tot percentages.

Het percentage meisjes in de groep wiskunde A/C is

$$\frac{\text{aantal meisjes met wiskunde A/C}}{\text{totaal aantal leerlingen met wiskunde A/C}} \times 100\%.$$

-  **9** **a** Bereken dit percentage met de gegevens van opgave 8.
b Bereken ook het percentage meisjes in de groep wiskunde B.
c Zet in de kruistabel op het werkblad per *wisgroep* de percentages jongens en meisjes in de cellen.
d Wat is je conclusie?

Als je het goed gedaan hebt, is het verschil in percentage meisjes in de A/C-groep en in de B-groep 20%. Dit verschil in percentage geeft een goede indicatie hoe sterk de verdeling op *geslacht* verschilt tussen beide groepen.

- 10** Stel dat de percentages meisjes in de ene groep en in de andere groep 0% verschillen.
a Wat is dan je conclusie over de samenhang van *geslacht* en *wiskunde-keuze*?

Stel dat de percentages meisjes in de ene groep en in de andere groep 100% verschillen.

- b** Wat is dan je conclusie over de samenhang van *geslacht* en *wiskunde-keuze*?

Als je verschillen constateert, moet je die vervolgens interpreteren. Dan krijg je met de vraag te maken of het geconstateerde verschil gering, redelijk, groot of extreem is. En dat is subjectief. In een situatie zoals in opgave 9 vinden statistici dat er vanaf bijvoorbeeld 30% sprake is van een (redelijk) groot verschil.

-  **11** **a** Wat is het verschil in percentage jongens in de A/C-groep en de B-groep?

Dat verschil in percentages jongens is precies gelijk aan het verschil in percentages meisjes. En dat is niet toevallig.

- b** Leg uit dat die verschillen in percentages altijd beslist gelijk zijn.

Hierboven heb je **verticaal gepercenteerd**. De kolomtotalen zijn op 100% gezet en die 100% is verdeeld over de cellen erboven. Je hebt de tabel *verticaal* gepercenteerd omdat je wilde weten wat het verschil was in de verdeling jongen-meisje in de A/C-groep en in de B-groep. De twee percentages vergelijk je vervolgens *horizontaal*: je neemt het verschil van de naast elkaar geplaatste percentages.

Je kunt ook **horizontaal percenteren**. Dan wil je weten hoe de verdeling wiskunde AC-wiskunde B is onder de meisjes en onder de jongens. Dan zet je de rijtotalen op 100% en vergelijkt vervolgens de gevonden percentages verticaal.

-  **12 a** Percenteer op het werkblad de tabel horizontaal.
b Wat is het percentageverschil?
 Is dit gelijk aan het percentageverschil bij verticaal percenteren?

Bij een kruistabel kun je de percentages dus op twee manieren uitrekenen en met elkaar vergelijken.

- Je kunt per kolom de percentages uitrekenen. Onderaan in de kolommen kom je dan op 100% uit. Deze percentages vergelijk je dan horizontaal met elkaar. Dit heet **verticaal percenteren, horizontaal vergelijken**.
- Je kunt ook per rij de percentages berekenen. Rechts aan het eind van de rijen krijg je 100%. Deze percentages vergelijk je dan verticaal met elkaar. Dit heet: **horizontaal percenteren, verticaal vergelijken**.

graduele verschillen

Wiskunde A en wiskunde C bereiden voor op een studie in de alfa- of gammavakken. Je kunt je dus afvragen of je tussen wiskunde A/C- en wiskunde B-leerlingen verschillen vindt in 'belangstelling voor kunst'. Die is gemeten met de variabele *kunstbel*. Deze variabele kan de waarden 1 tot en met 6 hebben (1 = geen belangstelling, 6 = veel belangstelling).

<i>leerling</i>	<i>geslacht</i>	<i>gebjaar</i>	<i>gebmd</i>	<i>gewicht</i>	<i>lengte</i>	<i>kunstbel</i>	<i>talen</i>	<i>cijfwis</i>	<i>cijfgem</i>	<i>huiswerk</i>	<i>wisgroep</i>	<i>profiel</i>	<i>leuk</i>	<i>tedoen</i>	<i>past</i>	<i>advies</i>
16	v	91	12	52	170	4	2	8	7,3	10	B	NG	3	1	2	4
17	v	92	1	59	165	3	3	8	7,0	15	A/C	EM	1	2	4	3
18	m	92	3	69	186	2	2	7	7,5	10	B	NG	4	1	4	4
19	m	91	10	75	190	2	2	6	6,5	7	A/C	EM	2	1	4	3
20	m	91	12	80	178	3	2	8	7,0	7	A/C	CM	2	1	4	3
21	v	92	6	61	177	3	2	8	7,4	6	B	NG	4	4	1	4
22	m	92	6	80	189	1	1	7	6,8	1	A/C	EM	3	1	2	4

-  **13** Bekijk de scores op *kunstbel*.
a Wie scoren het hoogst, de B-leerlingen of de A/C-leerlingen? Hoe heb je dat vastgesteld?

Als je de frequenties in een kruistabel zet, krijg je in dit geval een zogenaamde 6×2-tabel, want de variabele *kunstbel* heeft zes mogelijke waarden! Hieronder zie je de lege kruistabel:

<i>kunstbel</i>	<i>wisgroep</i>		totaal
	A/C	B	
1. geheel niet			
2. niet zo groot			
3. middelmatig			
4. tamelijk groot			
5. groot			
6. erg groot			
totaal	4	3	7

- b** Vul op het werkblad de frequenties in de kruistabel in.
- c** Vind jij dat de verdeling op *kunstbel* veel verschilt tussen A/C- en B-leerlingen? Waarom?

Als je voor alle 154 leerlingen de frequenties in een kruistabel zet, krijg je:

<i>kunstbel</i>	<i>wisgroep</i>		totaal
	A/C	B	
1. geheel niet	8	11	19
2. niet zo groot	11	26	37
3. middelmatig	11	44	55
4. tamelijk groot	9	17	26
5. groot	2	10	12
6. erg groot	2	3	5
totaal	43	111	154

- 14** Waarom is het lastig in de tabel de A/C-leerlingen en de B-leerlingen met elkaar te vergelijken wat kunstbelangstelling betreft?

Vanwege de ongelijke groepsgroottes kun je weer beter overgaan op relatieve frequenties.

-  **15 a** Vul op het werkblad de percentages in voor beide wiskundevakken. Rond af op hele percentages.
- b** In welk opzicht krijg je nu een beter inzicht in de verschillen?
- c** In welk opzicht is het vergelijken nog steeds lastig?

Begrepen?

- Een autobedrijf wil weten wat de klanten van het bedrijf vinden en stelt daarom een vragenlijst op.

Noem een aantal variabelen die daarbij een rol kunnen spelen?

- Een klein onderzoek naar drugsgebruik onder jongeren leverde de volgende gegevens op.

	gebruikt drugs	geen drugs	totaal
allochtoon	8	12	
autochtoon	4	20	
totaal			

Wat het verschil in percentage drugsgebruikers bij allochtonen en autochtonen?

Wat is het verschil in percentage allochtonen bij drugsgebruikers en geen-drugsgebruikers?

- We bekijken de resultaten van een tweede onderzoek naar drugsgebruik onder jongeren:

	gebruikt drugs	geen drugs	totaal
allochtoon	20	30	50
autochtoon			80
totaal			

Vul de tabel verder in als er geen verschil is in drugsgebruik tussen allochtone en autochtone jongeren.

Hoe ziet de tabel eruit als er een extreem verschil in drugsgebruik is tussen allochtone en autochtone jongeren?

1.2 Data in beeld

geclusterd staafdiagram, reepdiagram, cumulatief frequentiepolygoon

Bij de variabele *geslacht* zijn er maar twee waarden mogelijk: ‘man’ en ‘vrouw’. In dat geval kon je met het percentageverschil goed aangeven hoezeer beide wiskgroepen verschillen.

De variabele *kunstbel* heeft zes mogelijke waarden. Dan is het veel lastiger om het verschil tussen beide wiskgroepen te karakteriseren.

In de kruistabel staat *freq* voor **frequentie** (= aantal) en *p* voor percentage.

<i>kunstbel</i>	<i>wiskgroep</i>			
	<i>A/C</i>		<i>B</i>	
	<i>freq</i>	<i>p</i>	<i>freq</i>	<i>p</i>
1. geheel niet	8	18,6%	11	9,9%
2. niet zo groot	11	25,6%	26	23,4%
3. middelmatig	11	25,6%	44	39,6%
4. tamelijk groot	9	20,9%	17	15,3%
5. groot	2	4,7%	10	9,0%
6. erg groot	2	4,7%	3	2,7%
totaal	43	100,1%	111	99,9%

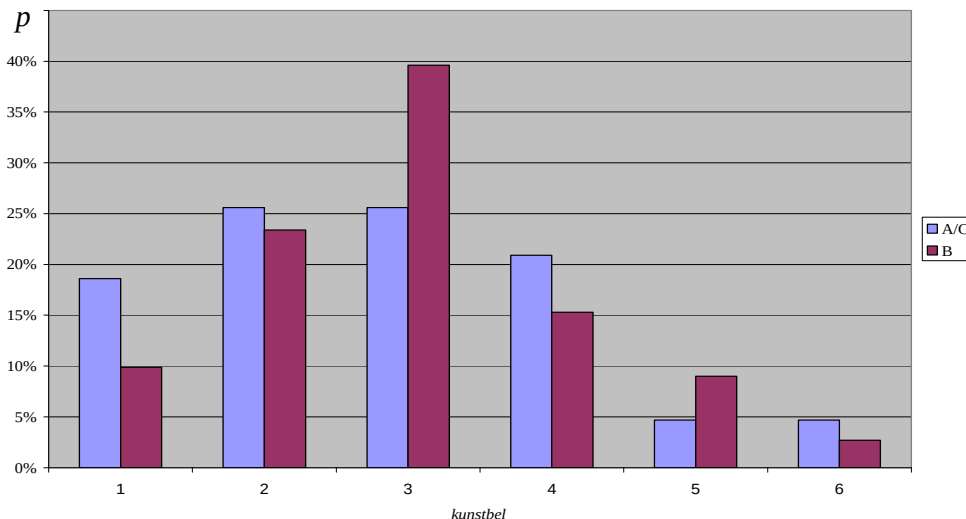
 **16** Het totale percentage A/C is niet precies 100%.

a Hoe komt dat?

Bij A/C is het totale percentage 0,1% te groot en bij B 0,1% te klein.

b Is dat toeval?

Met de percentages is het **geclusterde staafdiagram** hieronder gemaakt. Daarin staat hoe de belangstelling voor kunst verdeeld is over de verschillende waarden, van beide groepen, in procenten. Je krijgt meteen al een eerste indruk van de verschillen: de B-groep heeft een veel hoger percentage op de score *middelmatig* en ook een hoger percentage op de score *groot*; op de andere scores zijn de percentages lager dan van de A/C-groep.



 17 a Welk niveau van belangstelling komt het meeste voor als je de hele groep bekijkt?

In de grafiek zie je dat de aantallen niet symmetrisch zijn verdeeld rond 3,5. De aantallen links ervan zijn groter dan de aantallen rechts ervan.

b Wat betekent dat voor de algemene belangstelling voor kunst in deze groep leerlingen?

c Vind jij dat er veel verschil is in belangstelling voor kunst tussen beide groepen? Waarom wel/niet?

Om de groepen beter te kunnen vergelijken, kijken we naar de **cumulatieve percentages**. Je neemt dan de percentages tot en met een bepaalde score. Bijvoorbeeld het percentage A/C-leerlingen met een score t/m *middelmatig* is $18,6\% + 25,6\% + 25,6\% = 69,8\%$, afgerond 70%.

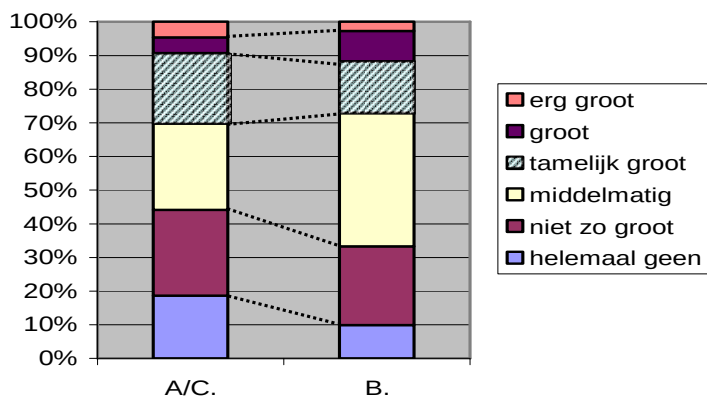
<i>kunstbel</i>	<i>wisgroep</i>					
	A/C			B		
	<i>freq</i>	<i>cf</i>	<i>cp</i>	<i>freq</i>	<i>cf</i>	<i>cp</i>
1. geheel niet	8	8	19%	11	11	10%
2. niet zo groot	11	19	44%	26	...	...%
3. middelmatig	11	30	70%	44	...	...%
4. tamelijk groot	9	39	91%	17	...	...%
5. groot	2	41	95%	10	...	...%
6. erg groot	2	43	100%	3	...	100%
totaal	43			111		

In de kolom *cf* staan de cumulatieve (= opgetelde) frequenties en *cp* staat voor cumulatief percentage. Zo is bij de A/C-groep de cumulatieve frequentie van “middelmatig” 30, want er zijn $8+11+11 = 30$ A/C-leerlingen met middelmatige of lagere belangstelling voor kunst, en dat is ongeveer 70%.

Cumulatief betekent *samenvoegend*, *ophopend*.

 18 Vul op het werkblad de ontbrekende cumulatieve percentages voor de B-groep in.

Door de staven in het staafdiagram van de vorige bladzijde te stapelen, krijg je automatisch de cumulatieve percentages in beeld. Dat zie je in het **reepdiagram**. De zes percentages van een groep (kolom *cp* in de kruistabel) zijn er opgestapeld weergegeven door (gekleurde) blokjes in een reep. Voor de duidelijkheid zijn de grenzen tussen de blokjes nog met stippellijnen verbonden.

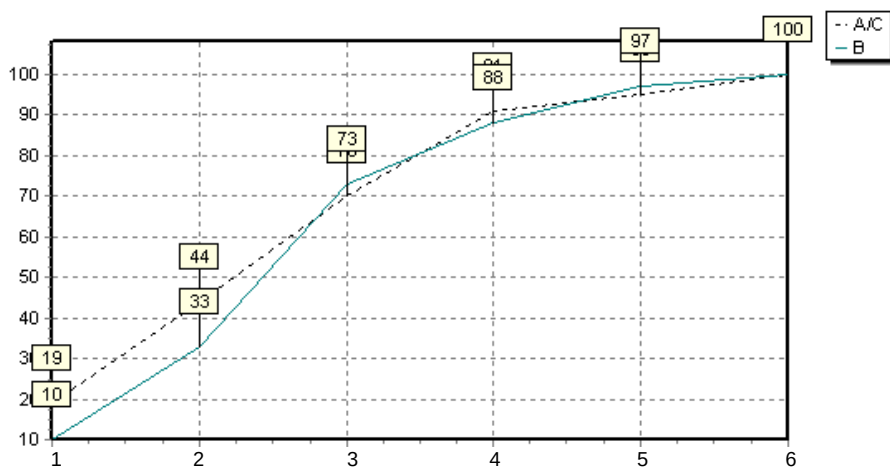


- 19 a Komt een lagere belangstelling voor kunst (dat is “helemaal geen” of “niet zo groot”) meer voor bij de B- dan bij de A/C-groep?
 b Wat is je conclusie over het verschil in *kunstbel* voor beide groepen?

Je kunt de verschillen tussen de cumulatieve percentages bekijken: $cp(A/C) - cp(B)$.
 Deze percentageverschillen noemen we V_{cp} .

- 20 a Vul op het werkblad de kolom V_{cp} in en stel vast wat het grootste absolute verschil is.
 b Hoe zie je in het reepdiagram wat het grootste verschil is?

We tekenen de lijngrafiek van de cumulatieve percentages cp van beide groepen.



- 21 a Lees uit de grafiek af hoeveel procent van de B-leerlingen middelmatig of minder belangstelling voor kunst hebben.
 b Lees uit de grafiek af hoeveel procent van de B-leerlingen tamelijk grote belangstelling voor kunst hebben (dat is categorie 4)?
 c Hoe zie je in de grafiek dat er procentueel meer A/C-leerlingen zijn met tamelijk grote belangstelling (categorie 3) voor kunst dan B-leerlingen?
 d Hoe zie je aan deze lijngrafiek dat de belangstelling voor kunst in de wiskunde B-groep (iets) groter is dan in de wiskunde A-groep?

Een lijngrafiek zoals hierboven heet een **cumulatieve frequentiepolygoon**.

Polygoon betekent letterlijk *veelhoek*. Je tekent die door de losse punten (hier bij de cumulatieve percentages) te verbinden. Je krijgt zodoende in het algemeen een grafiek met knikken.

De variabele *advies* zegt in hoeverre het advies van mensen belangrijk was voor een leerling om het gekozen profiel te kiezen. 1 betekent minst belangrijk, 4 betekent meest belangrijk. Hieronder staat de kruistabel van advies tegenover de wisgroep.

advies	wisgroep					
	A/C			B		
	freq	cf	cp	freq	cf	cp
1. niet belangrijk	3			10		
2. minder belangrijk	2			14		
3. belangrijk	14			25		
4. erg belangrijk	24			62		
totaal	43			111		

- 22 a Vul op het werkblad de tabel verder in.
b Teken de cumulatieve procentuele frequentiepolygonen voor beide wisgroepen in één figuur.
c Lees hieruit af voor welke groep het advies het zwaarst telde.

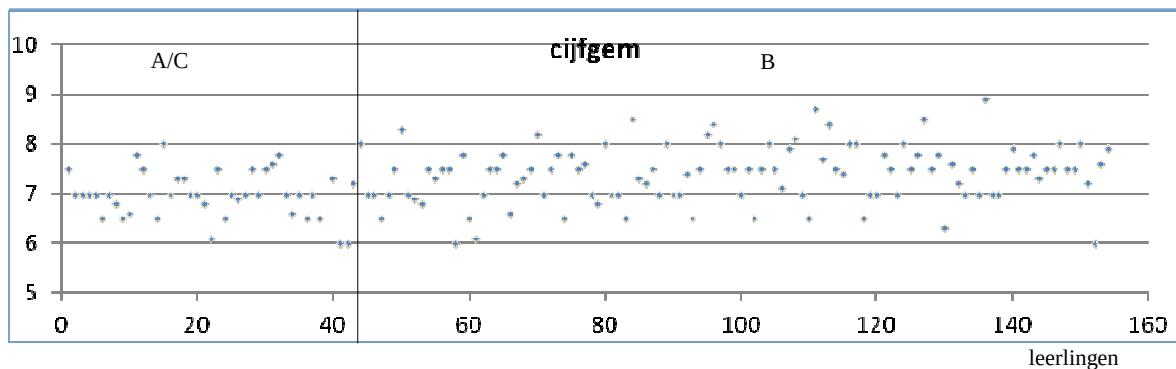
histogram, cumulatief frequentiepolygoon, boxplot

We gaan het gemiddelde rapportcijfer van de A/C- en de B-leerlingen vergelijken, bij de overgang van klas 3 naar klas 4. Dit gemiddelde is op één decimaal afgerond.

We gaan deze data op allerlei manieren weergeven:

- met een dotplot
- met een staafdiagram
- met een histogram
- met een cumulatieve frequentiepolygoon
- met een boxplot

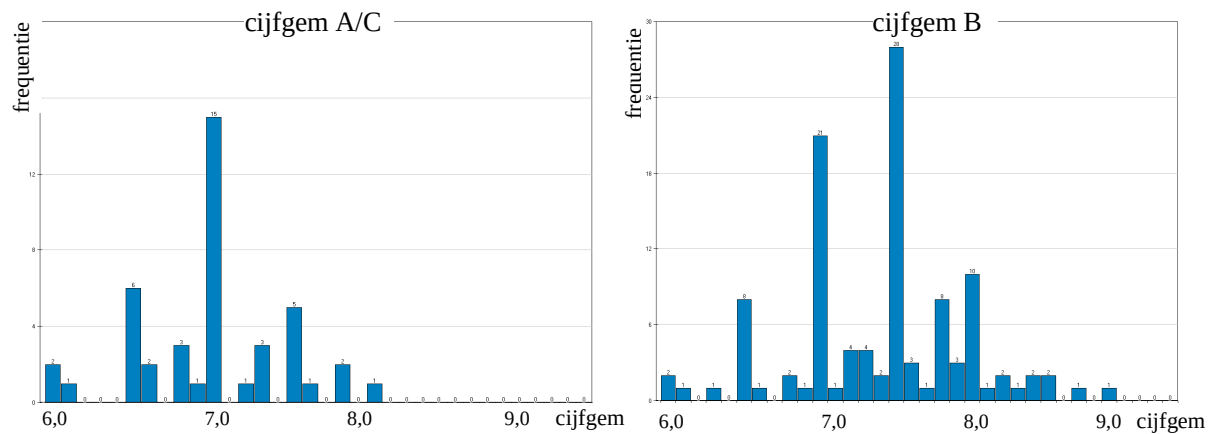
• Dotplots



De dotplot geeft alle data weer van beide groepen. Van alle leerlingen is het gemiddelde rapportcijfer aangegeven door een stip. Verticaal staan de cijfers, horizontaal de (genummerde) leerlingen, links A/C en rechts B.

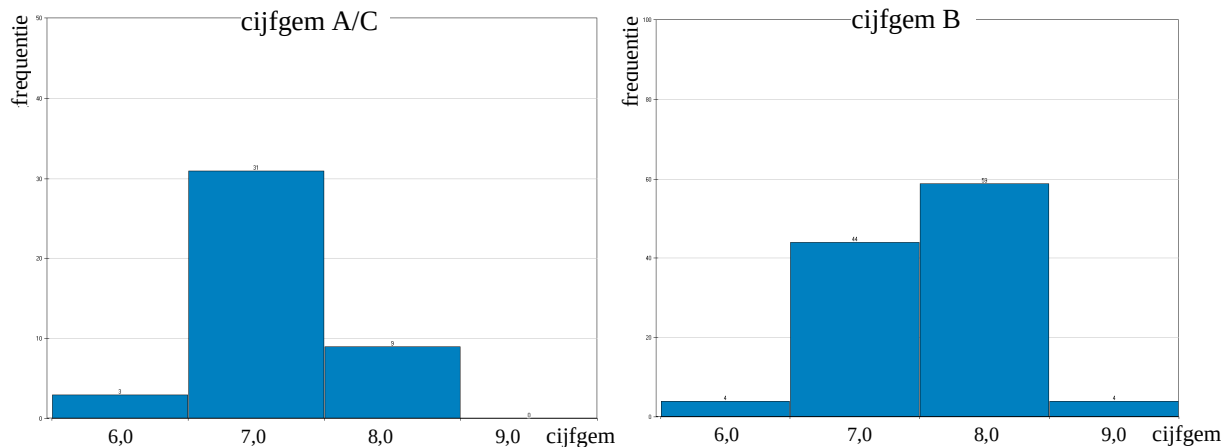
- 23 Welk verschil tussen beide groepen valt onmiddellijk op als je de dotplot van de A-groep vergelijkt met de dotplot van de B-groep?

- **Staafdiagrammen**



De hoogte van een staaf geeft aan hoe vaak de betreffende score voorkwam.
De staafdiagrammen maken je niet veel wijzer over de verschillen tussen beide groepen. Er ontstaat een overzichtelijker beeld als de gemiddelde eindcijfers worden ingedeeld in klassen.

- **Histogrammen**



24 Welk verschil tussen beide wisgroepen valt nu het meeste op?

De balk boven score 6 in het histogram voor de B-groep geeft weer dat 3 leerlingen een cijfer hadden tussen 5,5 en 6,5. Om precies te zijn: *minstens* een 5,5 en *lager* dan een 6,5. Die grenzen worden door de maatstreepjes op de horizontale as aangegeven. De score 7 is een **klasse** van scores, met **klassegrenzen** 6,5 en 7,5 en met **klassेमidden** 7.

Voordat je een histogram maakt, kies je een geschikte **klasse-indeling** van de scores.

Tip

Afhankelijk van de situatie zorg je ervoor dat er niet teveel (minder dan 10) en niet te weinig (ten minste 5) klassen zijn.

“Histogram” is afgeleid uit het Grieks: *histos* = balk, *graphein* = schrijven.

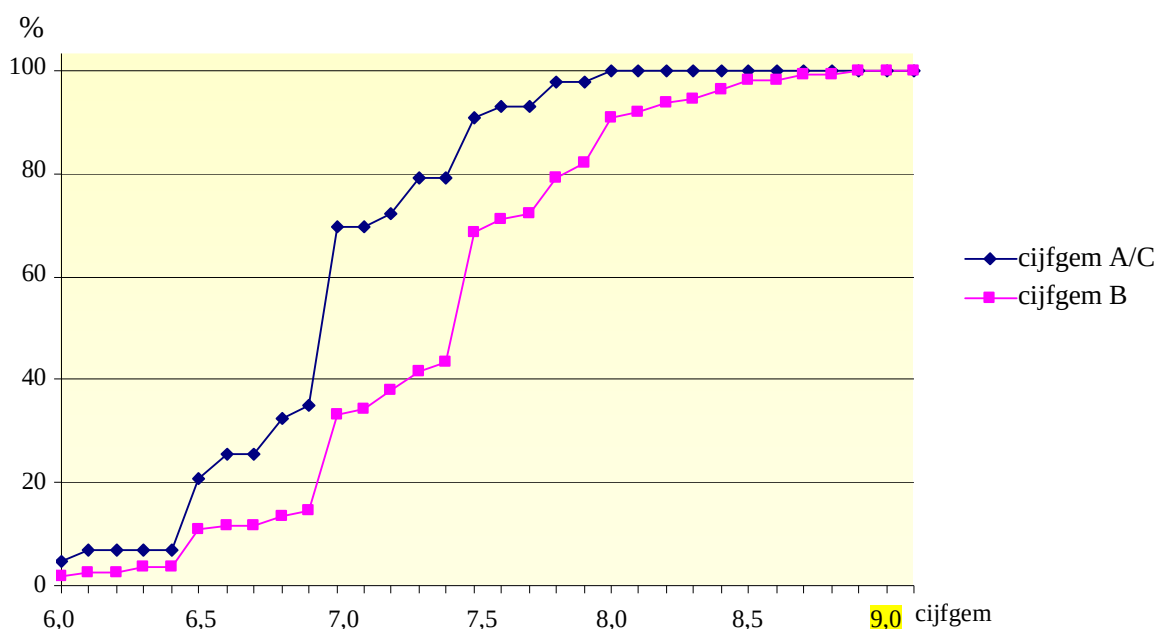
Opmerking

De begrippen staafdiagram en histogram lijken veel op elkaar. Omdat het gemiddelde cijfer in principe alle waarden kan aannemen tussen 1 en 10, heb je hier te maken met een **continue variabele**. In de praktijk gebruik je een histogram alleen bij continue variabelen. In alle andere gevallen maak je een staafdiagram met losse staven; de variabele neemt dan alleen losse waarden aan, zoals 7,5 ; 7,6 ; 7,7 ; enz.

- 25 Doordat elk cijfer eigenlijk een klasse is, worden er veel cijfers op één hoop gegooid. Welke informatie ben je kwijtgeraakt als je de histogrammen met de staafdiagrammen vergelijkt?

• Cumulatieve frequentiepolygonen

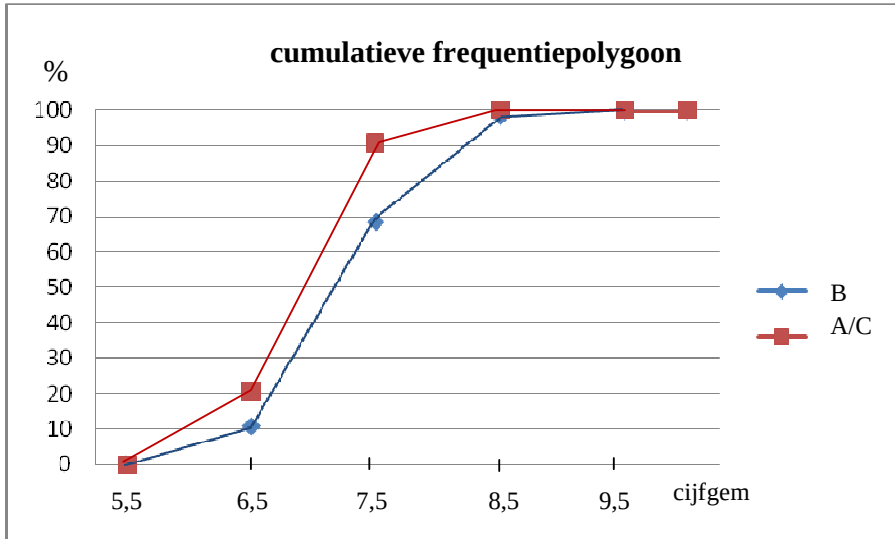
Een betere mogelijkheid om het verschil tussen de A/C-groep en de B-groep in beeld te brengen is de cumulatieve frequentiepolygoon of somfrequentiepolygoon. Op de horizontale as staan de gemiddelde cijfers, op de verticale as de cumulatieve percentages.



- 26 Bekijk de figuur hierboven.
- Bij welke waarde van *cijfgem* is het verschil tussen de cumulatieve percentages van de twee groepen het grootst?
 - Hoe groot is dat verschil?
 - Geef je mening over de verschillen tussen beide groepen wat het gemiddelde cijfer in vwo 3 betreft.

Je kunt ook een cumulatieve frequentiepolygoon maken door uit te gaan van de histogrammen (in plaats van van de waarden zelf).

Hieronder staan de cumulatieve frequentiepolygoon voor de A/C- en B-groep in één figuur. De stip boven 7,5 op de polygoon voor de B-groep zegt dat in de B-groep het percentage leerlingen met gemiddelde cijfer lager dan 7,5 bijna 70% is.



- 27 a. Lees uit de cumulatieve frequentiepolygoon af hoeveel procent van de A/C-leerlingen minstens 6,5 en hoogstens 7,4 als gemiddelde hadden.
 b. Lees dit percentage ook af uit de figuur op de vorige bladzijde.
 c. Hoe zie je in de figuur dat de gemiddelde rapportcijfers van de B-leerlingen hoger waren dan die van de A/C-leerlingen?

• Boxplots

Een andere manier om naar de gegevens te kijken is de volgende. Deel de populatie in vier gelijke stukken (kwarten). Zet daarvoor eerst alle data op volgorde van grootte. Als voorbeeld nemen we de gemiddelde cijfers in de A/C-groep. Dat zijn er 43:

6,0 ; 6,0 ; 6,1 ; 6,5 ; 6,5 ; 6,5 ; 6,5 ; 6,5 ; 6,5 ; 6,6 ; 6,6 ; 6,8 ; 6,8 ; 6,8 ; 6,9 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; **7,0** ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,0 ; 7,2 ; 7,3 ; 7,3 ; 7,3 ; 7,5 ; 7,5 ; 7,5 ; 7,5 ; 7,5 ; 7,6 ; 7,8 ; 7,8 ; 8,0

Deze verzameling cijfers verdeel je in twee gelijke helften van elk 21 cijfers. Het middelste cijfer is 7,0; dat is hierboven vet gemaakt. Links daarvan staan 50% van de data en rechts daarvan ook. Dit is de **mediaan** van de cijfers. Bij een even aantal cijfers zijn er twee middelsten. In dat geval nemen we het gemiddelde van die middelsten; dat is dan de mediaan.

De mediaan verdeelt de cijfers in twee even grote helften.

Op dezelfde manier bepaal je het middelste cijfer van de linkerhelft en ook van de rechterhelft. Die zijn onderstreept: 6,6 en 7,3. Dat zijn de zogenaamde kwartielen. 6,6 is het **eerste kwartiel**, 7,3 is het **derde kwartiel**.

Het tweede kwartiel is de mediaan. De kwartielen verdelen de verzameling van alle data dus in vier gelijke stukken; elk van (ruwweg) 25%.

Opmerking

Het gaat hier om een globale verdeling; kwartielen worden gebruikt voor grote databestanden. Dan doet het er niet toe of je de mediaan zelf wel of niet mee laat doen bij de kleinste helft om het eerste kwartiel te bepalen. Voorbeeld:

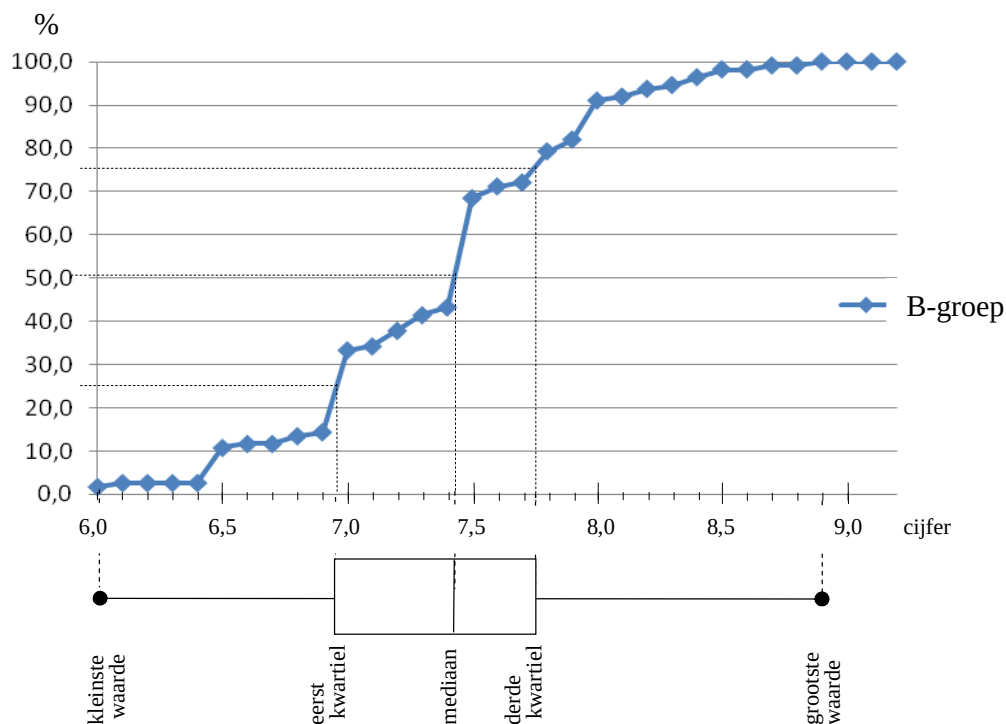
het eerste kwartiel van de lengtes van Nederlandse 18-jarige jongens is 178,4 cm; dat betekent dat 25% van de Nederlandse 18-jarige jongens korter dan 178,4 cm is en 75% langer dan 178,4 cm. Niemand is exact 178,4 cm; de precieze lengte van 178,4 cm is niet interessant.

Om misverstanden te voorkomen zullen we de volgende afspraak gebruiken:

- het eerste kwartiel is de mediaan van de linkerhelft, exclusief de eventuele middelste score,
- het derde kwartiel is de mediaan van de rechterhelft, exclusief de eventuele middelste score.

-  **28** Kijk naar de 25%-, 50%- en 75%-waarden op de verticale as van de A/C-polygoon op bladzijde 17 en zoek de bijbehorende cijfers op de horizontale as.
Welke cijfers zijn dat?

Hieronder zie je nog eens de cumulatieve frequentiepolygoon van gemiddelde rapportcijfers van de B-groep. Daaronder is de zogenaamde **boxplot** getekend.



Een **boxplot** wordt getekend door op schaal de uiterste waarden en de kwartielen aan te geven. In de “box” (de rechthoek tussen de kwartielen) zit 50% van de waarden. In dit geval:

- laagste waarde: 6,0
- eerste kwartiel: 6,95
(want de 25% lijn valt midden tussen de 6,9 en de 7,0)
- mediaan: 7,42
- derde kwartiel: 7,73
- hoogste waarde: 8,9

- 29 a Teken op het werkblad de boxplot voor de A/C-groep onder de boxplot voor de B-groep.
 b Beschrijf op grond van de boxplots het verschil tussen beide groepen met betrekking tot het gemiddelde eindcijfer in klas 3.
 c Heeft het eindcijfer voor wiskunde in klas 3 een rol gespeeld bij de keuze voor wiskunde A/C of wiskunde B? Geef argumenten.

Tussen het eerste en het derde kwartiel zit de helft van de scores. De afstand van het eerste en derde kwartiel heet **kwartielfstand**. Dat is de breedte van de box.

Als de kwartielfstand groot is, is de *spreiding* van de scores dat ook, dwz de scores liggen dan ver uit elkaar. (We komen hier later op terug.)

• Steelbladdiagram

Nauw verwant aan het histogram is het zogenaamde **steelbladdiagram**. We leggen dat uit aan de hand van de gewichten van de leerlingen in de A/C-groep.

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

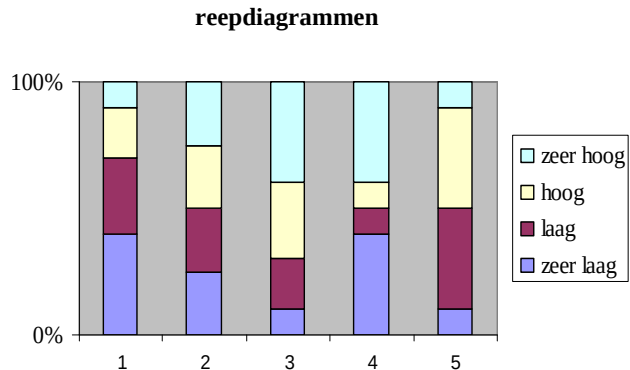
Op de bovenste regel staan de gewichten 41, 46 en 48 kg. De "4" staat in de "steel", de "1", "6" en "8" zijn de bladeren.

- 30 a Hoe maak je van het steelbladdiagram een histogram? Met welke klassebreedte?
 b Hoeveel A/C-leerlingen wegen meer dan 70 kg?
 c Bepaal de kwartielen en de mediaan op de manier van bladzijde 18.
 d Teken de cumulatieve frequentiepolygoon.
 e Bepaal de mediaan en de kwartielen uit de cumulatieve frequentiepolygoon.
 f Teken de boxplot.

Begrepen?

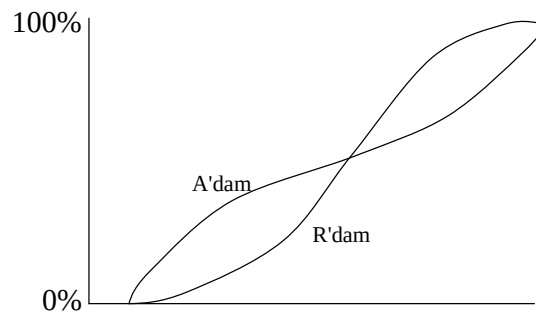
- Mensen is gevraagd naar hun interesse voor de aanstaande verkiezingen voor het Europese parlement. De antwoorden worden weergegeven in een reepdiagram. We bekijken vijf theoretisch mogelijke reepdiagrammen.

Zeg bij elk van deze reepdiagrammen in woorden wat het jou vertelt.



- De scores op een of andere variabele lopen van 120 tot 220. Ze worden verdeeld in tien klassen, met klassebreedte 10. Veronderstel dat de scores *uniform verdeeld* zijn, dat wil zeggen dat alle klassen evenveel scores tellen.
 Wat zijn dan de kwartielen?
 Wat is de kwartielafstand?
 Hoe ziet het steelbladdiagram eruit, als de steel bestaat uit "12", "13", ..., "22"
 Teken de boxplot.
 Hoe ziet het reepdiagram eruit?
 Hoe ziet de dotplot eruit?
 Hoe ziet de cumulatieve frequentiepolygoon eruit?

- Bekijk de twee cumulatieve frequentiepolygoon, van Amsterdammers en van Rotterdammers op een of andere variabele.
 Hoe kun je zien welke stad de meeste spreiding in de antwoorden heeft?
 Schets de boxplot van elk van de steden.



1.3 Computerpracticum

Bij grote databestanden is het ondoenlijk om met de hand het tel- en tekenwerk te doen. Met een statistiekprogramma neemt de computer je al dit werk uit handen. Er bestaan veel statistiek-programma's. Ze hebben allemaal hun voordeel en hun nadeel. We gebruiken in dit practicum het programma VUStat om de gegevens van de 154 vwo4-leerlingen te analyseren.

- Dubbelklik in Digimap op Verschillen en vervolgens op Copmputerpracticum.

Je krijgt een datatabel (datamatrix).
Hiernaast zie je het linkerbovengedeelte van de tabel.

leerling	geslacht	gebjjaar	gebmd	gewicht	lengte	kunstbel
1	m	92	april	70	190	middelmatig
2	m	92	mei	55	170	niet zo groot
3	m	92	januari	60	170	niet zo groot
4	v	91	november	52	160	niet zo groot
5	v	92	september	60	175	middelmatig
6	v	92	januari	65	178	middelmatig
7	v	92	januari	55	169	middelmatig
8	v	91	mei	52	168	groot
9	v	91	oktober	52	168	tamelijk groot
10	v	92	september	64	175	tamelijk groot

- We gaan een **kruistabel** maken waarbij het geslacht wordt uitgezet tegen het profiel.
- Kies in het menu Tabel > Kruistabel
- Selecteer de variabele *geslacht* en klik bij Rijvariabele op ; selecteer zo ook de kolomvariabele *profiel*.

Je krijgt het volgende resultaat:

geslacht	profiel				Totaal
	CM	EM	NG	NT	
jongen	2	15	17	35	69
meisje	9	27	43	6	85
Totaal	11	42	60	41	154

VUStat kan de aantallen in de kruistabel ook omzetten in percentages.

- Kies rij-percentages.
- Kies kolom-percentages.
- We gaan de verdeling van de jongens over de vier profielen grafisch weergeven. We beginnen met het **staafdiagram en reepdiagram**.
- Kies in het menu Grafiek>Staafdiagram, selecteer de variabele *geslacht*.
- Voor een opdeling in profielen klik je op de knop Spliten. Selecteer de splitsvariabele *profiel*.
- Kies het (reep)diagram  en vink Procenten aan.
- Doe hetzelfde te beginnen met *profiel* en als splitsvariabele *geslacht*.

► Maak een geclusterd diagram met de knop 

NB. Met de knop Splitsing opheffen kun je een splitsing ongedaan maken.

- We gaan een **cirkeldiagram** maken

In een cirkeldiagram (of sectordiagram) komt de grootte van elk cirkelsegment overeen met de relatieve frequentie van een score.

Je gaat een cirkeldiagram maken bij de profielkeuze van de jongens.

- Kies in het menu Grafiek > Cirkeldiagram en selecteer variabele *profiel*.
Voor een groepsindeling splits je op de variabele *geslacht*.
Als Bijschrift is hier Label&Percentage geschikt.

We maken ook een cirkeldiagram voor de verdeling van de meisjes over de profielen. Daarvoor moeten we eerst de meisjes uit het hele bestand selecteren.

- Kies Data > Selectiefilter > Filter ; selecteer de variabele *geslacht* en dan label v.
- Kies Grafiek > Cirkeldiagram enz.

NB. Via Data kun je de selectie opheffen.

- We gaan een **histogram** maken van de lengten van de leerlingen.

- Kies in het menu Grafiek > Staafdiagram en selecteer de variabele *lengte*.

Je krijgt meer inzicht in de verdeling van de variabele *lengte* als je ook andere klasse-indelingen bekijkt.

- Klik op Indeling en zet het aantal klassen op 3; kies een klassebreedte.
- Bekijk zo verschillende indelingen.

- We gaan een **cumulatieve frequentiepolygoon** maken van de kunstbelangstelling van de leerlingen.

- Kies in het menu Grafiek > Lijndiagram en selecteer de variabele *kunstbel*.
- Vink Cumulatief aan, en eventueel Procenten.

- We gaan een **boxplot** maken.

Je gaat een boxplot maken van het gewicht van de leerlingen.

- Kies in het menu Grafiek > Boxplot en selecteer de variabele *gewicht*.

Klik op  en dan op OK.

1.4 Onderzoek



Het bestand sportprestaties.vus (voor VU-Statistiek) bevat gegevens van 74 leerlingen over de drie sporten Sprint, Verspringen en Vergooien (met een gewicht van 200 gram).

De onderzoeksvraag is:

Hoe verschillen jongens en meisjes op elk van deze drie onderdelen en wat is het effect van de leeftijd op de prestaties?

Onderzoek de data op deze vraag met behulp van VUStat. Breng de verschillen in beeld en geef aan hoe groot deze verschillen zijn.

Verwerk in je conclusie in ieder geval de volgende aandachtspunten:

- beargumenteer je keuze van de diagrammen: waarom heb je juist die diagrammen gekozen?
- interpreteer de grootte van de waargenomen verschillen: vind je ze groot of klein? Waarom?

Spiekbriefje voor VU-Statistiek

► Dubbelklik in Digimap op Onderzoek 1

- **Kruistabel**

Kies in het menu Tabel > Kruistabel.

Selecteer een rijvariabele en een kolomvariabele.

Kies rij-percentages of kolom-percentages en sluit af met OK.

- **Staafdiagram / Reepdiagram**

Kies in het menu Grafiek>Staafdiagram, selecteer een variabele.

Klik op de knop Splitsen.

Selecteer een splitsvariabele.

Kies het (reep)diagram  en vink Procenten aan.

Kies geclusterd diagram .

- **Cirkeldiagram**

Kies in het menu Grafiek > Cirkeldiagram en selecteer een variabele.

Kies een Bijschrift.

- **Histogram**

Kies in het menu Grafiek > Staafdiagram en selecteer een continue variabele (blz. 17).

Klik op Indelingen en kies het aantal klassen en de klassebreedte.

- **Frequentiepolygoon**

Kies in het menu Grafiek > Lijndiagram en selecteer een continue variabele (blz. 17).

Vink Cumulatief aan voor een cumulatieve frequentiepolygoon.

- **Boxplot**

Kies in het menu Grafiek > Boxplot en selecteer een continue variabele (blz. 17).

- Probeer eens uit te zoeken of leerling 24 een jongen of een meisje is.

1.5 Samenvatting par. 1.0 t/m 1.4

Statistiek begint met het stellen van een vraag, bijvoorbeeld naar de verschillen tussen twee groepen leerlingen. Om die vraag te kunnen beantwoorden worden gegevens verzameld, bijvoorbeeld door middel van een enquête. De verzamelde gegevens worden gerangschikt in een **datamatrix**. In de verticale kolommen van de matrix staan de **scores** van de gemeten **variabelen**, in de horizontale rijen staan de **objecten** van het onderzoek.

Vaak is de datamatrix zo uitgebreid, dat je daaruit niet rechtstreeks conclusies kunt trekken. Daarom worden de data gerepresenteerd in getallen, kleinere tabellen of in diagrammen.

In getallen: **mediaan, kwartielen**

In kleinere tabellen: **frequentietabel, kruistabel**

In diagrammen **cirkeldiagram, staafdiagram, histogram, frequentiepolygoon, boxplot**.

De **mediaan** is de waarde die de op volgorde gezette scores van een variabele in twee helften verdeelt: 50% van de scores heeft een waarde die kleiner is dan de mediaan en 50% een waarde die groter is dan de mediaan. De **kwartielen** zijn de waarden die de geordende scores in vier opeenvolgende kwarten verdelen.

Bij een oneven aantal is de mediaan de middelste waarneming, bij een even aantal het gemiddelde van de middelste twee waarnemingen.

In een **frequentietabel** staat hoe vaak de verschillende scores op een variabele voorkomen.

De **relatieve frequentie** van een score is het percentage van het totaal waarin de score voorkomt.

In formule: $\text{relatieve frequentie (in \%)} = \frac{\text{frequentie}}{\text{totale aantal}} (\times 100\%)$

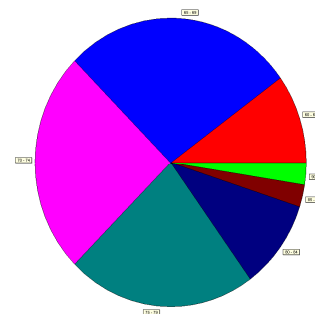
Een **cumulatieve (relatieve) frequentie** van een score is de (relatieve) frequentie van die score plus die van alle lagere waarden (in procenten).

In een **kruistabel** worden twee variabelen tegen elkaar uitgezet. Er staan de frequenties of de relatieve frequenties in waarin de combinaties van scores voorkomen. Voorbeeld:

geslacht	wisgroep		totaal
	A/C	B	
man	13	56	69
vrouw	30	55	85
totaal	43	111	154

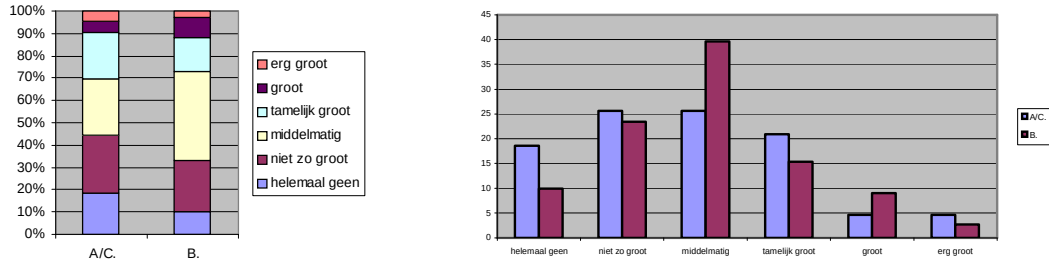
Een **steelbladdiagram** geeft een overzicht van alle scores, gerangschikt in een verticale “steel” en horizontale “bladeren”. In een **cirkeldiagram** worden de percentages van de scores uitgezet als sectoren in een cirkel. Voorbeeld:

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



In een **staafdiagram** worden de (relatieve) frequenties van de scores uitgezet als losse staven. Een **histogram** is een staafdiagram bij een continue variabele (de scores kunnen in principe *alle* getallen in een domein aannemen). De scores zijn ingedeeld in klassen. De staven staan aan elkaar vast.

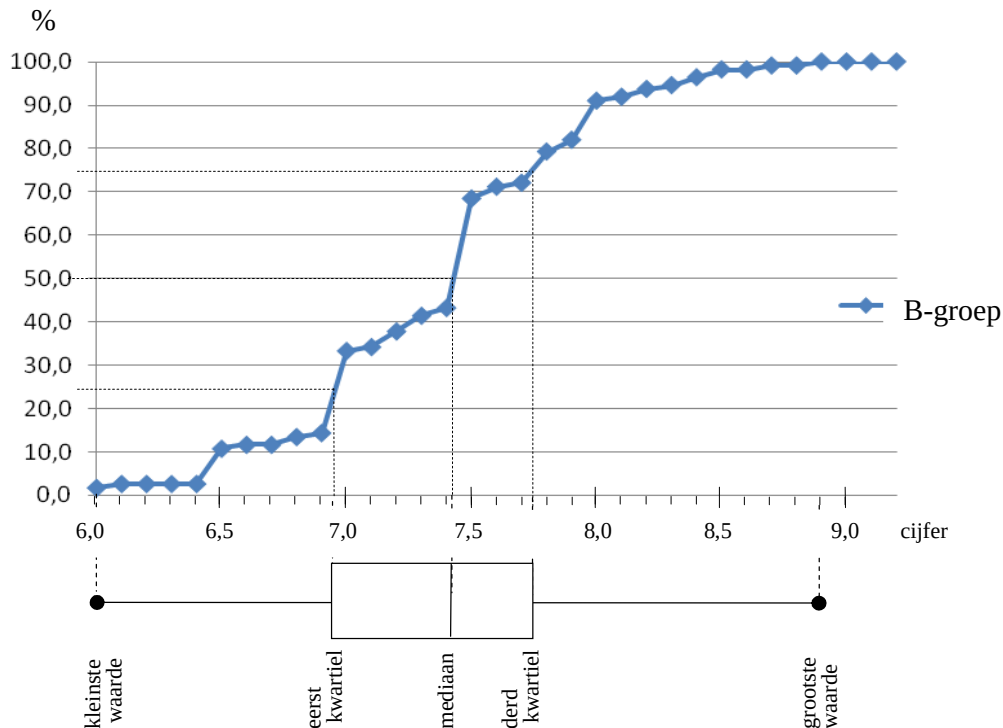
Voorbeeld van een **reepdiagram** (links) en een **geclusterd staafdiagram** (rechts):



Een **frequentiepolygoon** is een lijndiagram bij een continue variabele die de (relatieve) frequenties met elkaar verbindt.

Bij een **cumulatieve frequentiepolygoon** zijn de (relatieve) cumulatieve frequenties in een lijngrafiek gezet. Daaruit kun je bij een waarde aflezen hoeveel procent van de data een kleinere of gelijke waarde heeft.

Als je de uiterste waarden, de mediaan en het eerste en derde kwartiel kent, kun je de **boxplot** maken. Voorbeeld:



1.6 Verschillen meten

Temperatuur meet je met een thermometer. Of de een het nu wat aan de kille kant vindt en de ander juist een lekker temperatuurtje, de thermometer is onverbiddelijk. De thermometer geeft een getalswaarde aan de temperatuur die los staat van je persoonlijke mening.

Iets soortgelijks doet zich voor in de statistiek (en in veel meer gebieden van de wetenschap). In het voorgaande hebben we naar verschillen gekeken tussen twee groepen. Over de grootte van die verschillen kun je twisten. De een vindt dat ze aanzienlijk zijn, een ander vindt dat dat wel meevalt. We gaan nu proberen de verschillen te meten, dat wil zeggen uit te drukken in een getal. Welke manier van meten je kiest, hangt af van de situatie en soms van je persoonlijke voorkeur. Net zoals bij de thermometer: de Amerikanen doen het met Fahrenheit en in Europa houden we het op Celsius.

Maximale percentageverschil



Je zou kunnen verwachten dat leerlingen die voor wiskunde A/C gekozen hebben, meer belangstelling hebben voor kunst dan wiskunde B-leerlingen. De variabele *kunstbel* van de leerling-enquête geeft hier informatie over. De volgende kruistabellen geven de resultaten van het denkbeeldige Amalia College.

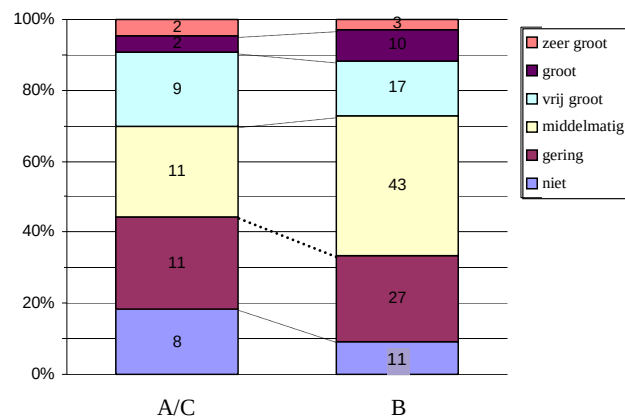
In de linker tabel staan de frequenties, in de rechter tabel de cumulatieve aantallen.

<i>kunstbel</i>	A/C	B
niet	8	11
gering	11	27
middelmatig	11	43
vrij groot	9	17
groot	2	10
zeer groot	2	3
	43	111

<i>kunstbel</i>	A/C	B
niet	8	11
gering	19	38
middelmatig	30	81
vrij groot	39	98
groot	41	108
zeer groot	43	111

Vervolgens kijken we naar de cumulatieve percentages en hun verschillen *Vcp*. Die staan in onderstaande tabel. In het reepdiagram daarnaast zijn die percentages in beeld gebracht.

<i>kbel cum.</i>	A/C	B	<i>Vcp</i>
niet	19%	10%	9%
gering	44%	34%	10%
middelmatig	70%	73%	3%
vrij groot	91%	88%	3%
groot	95%	97%	2%
zeer groot	100%	100%	0%



Je zoekt nu de maximale waarde van V_{cp} op. Dat is een geschikte maat voor het verschil tussen beide groepen op de variabele *kunstbel*. Hij heet het **maximale cumulatieve percentageverschil**. In dit geval is $\max. V_{cp} = 10\%$. Deze waarde zie je bij de reepdiagrammen terug als de meest steile verbindingslijn tussen de beide staven.

In stappen bepaal je het $\max. cp$ -verschil dus als volgt:

- Bepaal de cumulatieve frequenties van beide groepen.
- Zet die om in de cumulatieve percentages.
- Bepaal per waarde het (absolute) verschil van de cumulatieve percentages tussen de groepen.
- Neem van die verschillen de grootste waarde; dat is $\max. V_{cp}$.

In de praktijk hanteert men de regel:

- als $\max. V_{cp}$ kleiner dan 15% is, is het verschil gering,
- als $\max. V_{cp}$ tussen 15% en 30% is, is het verschil middelmatig,
- als $\max. V_{cp}$ groter dan 30% is, is het verschil groot.

- 31 Hieronder zie je de kruistabel van *aantal vreemde talen* en *wiskundekeuze* (A/C of B).
Bereken op het werkblad voor dit geval $\max. V_{cp}$. Is het verschil tussen beide groepen groot?

<i>aantal vreemde talen</i>	<i>wiskundegroep</i>						<i>Vcp</i>
	<i>A/C</i>			<i>B</i>			
	<i>freq</i>	<i>cf</i>	<i>cp</i>	<i>freq</i>	<i>cf</i>	<i>cp</i>	
een	7			16			
twee	26			58			
drie	10			30			
vier	0			7			
totaal	43			111			

- 32 Bij de hoogste waarde van de variabele is V_{cp} altijd 0. Leg dat uit.
- 33 Iemand overweegt bij de variabele *profiel* ook $\max. V_{cp}$ te berekenen om de jongens en de meisjes te vergelijken.
Waarom is dat niet zinvol?
- 34 Je zou als verschilmaat ook het gemiddelde van alle verschillen V_{cp} kunnen kiezen.
Bedenk een voor- en een nadeel van deze keuze ten opzichte van $\max. V_{cp}$.

- 35 In de kruistabel hiernaast, weer van het Amalia College, is het *geslacht* tegen de *wiskundekeuze* uitgezet.

<i>geslacht</i>	<i>A/C</i>	<i>B</i>	<i>tot</i>
m	13	56	69
v	30	55	85
totaal	43	111	154

- a Neem de tabel over en percenteer die verticaal.
- b Waarom is het in dit geval niet zinvol om $\max. V_{cp}$ uit te rekenen?
- c Nu $\max. V_{cp}$ als maat is afgefallen om het verschil in wiskundekeuze bij jongens en meisjes uit te drukken, moet je een andere maat bedenken. Welke maat bedenk jij?

Odds-ratio

In de laatste opgave van de vorige paragraaf was het niet zinvol om $\max.V_{cp}$ te bepalen, omdat er geen natuurlijke volgorde is tussen de twee geslachten man en vrouw. We zeggen dat *geslacht* een **nominale** variabele is (dat komt van het Latijnse woord *nomen*, wat *naam* betekent). Om $\max.V_{cp}$ zinvol te laten zijn moeten de waarden van de variabele een (natuurlijke) volgorde hebben. We spreken dan van een **ordinale** variabele (dat komt van het Latijnse woord *ordo*, wat *volgorde* betekent).

In het geval van de laatste opgave moeten we een andere maat verzinnen om grip te krijgen op de verschil man/vrouw in de A/C-groep en de B-groep.

Daarvoor vergelijken we de verhouding *aantal man* : *aantal vrouw* in beide groepen. Zo'n verhouding noemt men een **odd** ¹⁾.

Voor de B-groep is die gelijk aan $56 : 55 \approx 1,02$.

- 36 a Hoe groot is die verhouding bij de A/C-groep?
b Vind jij dat die verhoudingen in de A/C-groep en in de B-groep veel verschillen?

Het quotiënt van deze twee man : vrouw-verhoudingen heet de **ongelijkheidsverhouding** of de **odds-ratio** van de man : vrouw-verhouding tussen de A/C- en de B-groep.

c Laat zien dat de odds-ratio hier 2,37 is.

- 37 a Bereken de odds-ratio vanuit de tabel in opgave 34, maar nu voor de verhouding B : A/C tussen de vrouwen en de mannen.
b Verschilt deze waarde van de odds-ratio in opgave 35? Leg uit dat dat geen toeval is.
- 38 a Wat is de odds-ratio als er geen verschil is in de wiskundekeuze tussen mannen en vrouwen?
b Wat kun je over de odds-ratio zeggen als de wiskundekeuzes tussen mannen en vrouwen erg veel verschillen?

Let op de volgorde. Als je de waarde 2,44 hebt gevonden voor de man : vrouw-verhouding tussen twee groepen, is de odds-ratio voor de omgekeerde verhouding vrouw : man ook de omgekeerde waarde: $\frac{1}{2,44} = 0,41$.

Meestal wordt de volgorde zo gekozen dat de odds-ratio groter dan of gelijk aan 1 is. In de rest van het hoofdstuk zullen we ons aan deze afspraak houden.

Als waardering van de odds-ratio is gangbaar:

- als $\text{odds-ratio} < 2$, is het verschil gering,
- als $2 < \text{odds-ratio} < 3$, is het verschil middelmatig,
- als $\text{odds-ratio} > 3$, is het verschil groot.

¹⁾ Odds is een Engelse term die vooral bij weddenschappen wordt gebruikt. In een gokkantoor kan men wedden op de uitslag van een wedstrijd bijvoorbeeld van A tegen B. De **odds** van A is de wedverhouding "kans dat A wint" : "kans dat A verliest". De odds bepalen de uitbetaling die het kantoor doet als A wint en als A verliest.

- 39 Nevenstaande kruistabel geeft nog eens de verdeling van het aantal vreemde talen tegenover wiskundeB of -A/C.

Hoe groot is de odds-ratio voor een-of-twee talen : drie-of-vier talen tussen de twee groepen?

aantal vreemde talen	wiskundegroep		totaal
	A/C	B	
een of twee	33	74	107
drie of vier	10	37	47
totaal	43	111	154

- 40 a Maak een kruistabel voor de kunstzinnige belangstelling tegenover de wiskundegroep (zie bladzijde 10):
b Bereken de odds-ratio voor laag : hoog tussen de twee wiskundegroepen.

kunstzinnige belangstelling	wiskundegroep		totaal
	A/C	B	
laag (1,2,3)			
hoog (4,5,6)			
totaal			

- 41 In onderstaande kruistabel zijn de eindcijfers voor wiskunde in klas 3 (variabele *cijfwis*) uitgezet tegen de wiskundekeuze. Voor de A/C-groep zijn de frequenties *freq*, de cumulatieve frequenties *cf* en de cumulatieve percentages *cp* al ingevuld.

cijfwis	wiskundekeuze						Vcp
	A/C			B			
	freq	cf	cp	freq	cf	cp	
5	4	4	9,3%	0		 %	 %
6	12	16	37,2%	9			
7	15	31	71,1%	36			
8	11	42	97,7%	44			
9	0	42	97,7%	21			
10	1	43	100%	1			
totaal	43			111			

- a Vul op het werkblad zelf de tabel in voor de B-groep.
b Bepaal de waarde van *max.Vcp*.
c Waarom kun je hier de odds-ratio niet gebruiken?

Je kunt nu net zo'n truc uithalen als in opgave 39: neem de eindcijfers voor wiskunde samen in twee groepen: laag (5,6,7) en hoog (8,9,10).

- d Maak een bijbehorende kruistabel.
e Bereken de odds-ratio voor de verhouding laag : hoog tussen beide wiskundegroepen.

-  42 Waarop verschillen de A/C- en B-leerlingen van het Amalia College het meest: op geslacht, op aantal vreemde talen, op kunstzinnige belangstelling of op eindcijfer voor wiskunde in klas 3?

Begrepen?

- In een enquête is mensen met een hoog inkomen en mensen met een laag inkomen gevraagd of ze jaarlijks veel of weinig aan liefdadigheid gaven.

Is het mogelijk dat $\max.V_{cp}$ van de verdeling weinig/veel liefdadigheid in de twee inkomensgroepen 0% is?

Is het mogelijk dat $\max.V_{cp}$ van de verdeling weinig/veel liefdadigheid in de twee inkomensgroepen 100% is?

- In een enquête is aan mannen en vrouwen gevraagd of ze voor of tegen een zekere maatregel van de overheid zijn. Van de resultaten is een kruistabel gemaakt.

	voor	tegen
man	a	b
vrouw	c	d

Wat is de odds-ratio van de verhouding *voor* : *tegen* tussen de mannen en de vrouwen? Schrijf je antwoord zo eenvoudig mogelijk.

Wat is de odds-ratio van de verhouding *man* : *vrouw* tussen de voor- en tegenstemmers? Schrijf je antwoord zo eenvoudig mogelijk.

1.7 Gemiddelde en standaarddeviatie

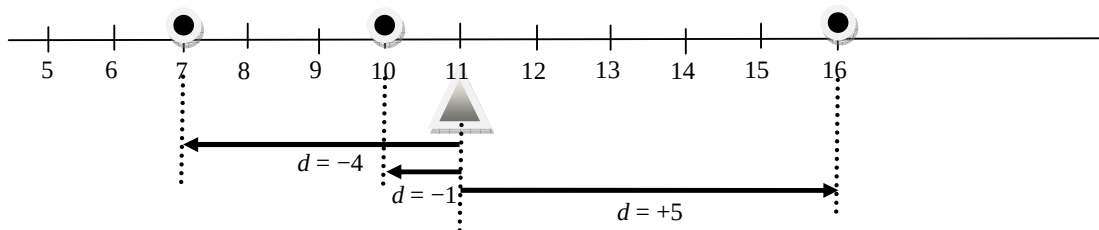
Naast het maximale cumulatieve percentageverschil $\max.Vcp$ en de odds-ratio zijn er nog andere methodes om naar verschillen te kijken. In deze paragraaf leer je zo'n methode aan de hand van het verschil tussen de A/C- en de B-groep in de variabele *huiswerk*: dat is de tijd die de leerlingen wekelijks besteden aan het maken van huiswerk.

43 Hieronder zie je het eerste stukje van de datamatrix van het Amalia College.

nummer	geslacht	gebjaar	gebmond	maanden	gewicht	lengte	kunstsbel	talen	cijfwis	cijfgem	huiswerk	wisvak	wisgroep	profiel	leuk	tedoen	past	advies
1	m	91	10	195	73	188	2	2	8	7,4	10	B	B	NG	2	4	1	3
2	v	92	11	182	62	178	5	3	6	7,1	6	C	A/C	CM	2	1	4	3
3	v	91	7	198	59	174	3	2	7	6,8	14	A	A/C	EM	1	2	3	4
4	v	91	12	193	54	176	4	2	9	7,8	7	B	B	NT	2	1	3	4
5	v	92	4	189	72	164	6	3	5	6,9	9	C	A/C	CM	1	2	4	3
6	m	92	2	191	73	179	3	2	7	6,5	16	B	B	NG	3	4	2	1
7	m	91	10	195	79	189	4	3	8	7,2	7	A	A/C	NG	4	3	1	2

- Kijk naar de drie leerlingen met wiskunde B. Bereken voor deze drie het gemiddeld aantal uren dat ze wekelijks aan huiswerk besteden.
- Doe dat ook voor de vier leerlingen met wiskunde A/C.

Hieronder zijn de scores 7, 10 en 16 van de drie B-leerlingen op een getallenlijn door balletjes weergegeven en de plaats van het **gemiddelde** met een wigje ▲. De drie scores wijken respectievelijk -4, -1 en +5 af van het gemiddelde. Dat is in het plaatje met pijlen aangegeven. De afwijkingen noemt men **deviaties**; daarvoor gebruikt men de letter d (van *deviatie* = *afwijking*). De drie d -waarden zijn hier respectievelijk -4, -1 en +5.

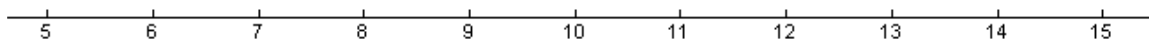


- 44 Met $\sum d$ wordt de som (optelling) van de drie deviaties bedoeld.
- Hoe groot is $\sum d$?

Het gemiddelde wordt wel de *evenwichtswaarde* van de drie scores genoemd.

- Waarom is dat een toepasselijke naam?

45 Hieronder zie je een stuk getallenlijn.



- Neem de figuur over en geef met balletjes de plaatsen aan van de vier scores uit de A/C-groep.
- Geef vervolgens de plaats van hun gemiddelde aan met een wigje ▲.

- c Geef de vier deviaties aan met pijlen en zet de d -waarde erbij.
 d Hoe groot is $\sum d$?

Met de getallenlijn als een wip (balans) kun je het gemiddelde beschouwen als het evenwichtspunt. De scores zijn even zware poppetjes op de wip. De deviaties links wegen op tegen de deviaties rechts: de som van de deviaties is 0.

Het gemiddelde is, net als de mediaan, een **centrummaat**; dat is een maat voor het centrum van de verdeling van de scores.

Het gemiddelde wordt aangegeven met een liggend streepje boven de gebruikte letter. Als je de letter h gebruikt voor het aantal huiswerkuren, is $\bar{h}$ hun gemiddelde.

Het gemiddelde bereken je door alle scores bij elkaar op te tellen en te delen door het totale aantal scores, in formule:

$$\bar{h} = \frac{\sum s}{n},$$

waarbij n het totale aantal scores is. Lees de teller als: "de som van de scores s ".

Extra

Stel we hebben n scores, die we s noemen. De gemiddelde score $\bar{s}$ is die waarde ten opzichte waarvan de som van de afwijkingen 0 is. Hiervan uitgaande vind je $\bar{s}$ als volgt. De afwijkingen zijn $s - \bar{s}$. De som daarvan is $\sum s - n \cdot \bar{s}$. Omdat dit gelijk aan 0 is, is $\bar{s} = \sum s / n$.

Soms komt een score vaker dan één keer voor. Het score 5 kan bijvoorbeeld 10 keer voorkomen. De score 5 heeft dan dus frequentie 10 en telt 10 keer mee in de berekening van het gemiddelde. In de som van de scores krijg je dus $\dots + 5 + 5 + 5 + 5 + 5 + 5 + 5 + 5 + 5 + 5 + \dots$. In plaats daarvan schrijf je liever $10 \cdot 5$, dus de *frequentie* maal de *score*.

Als s alle *verschillende* scores zijn met bijbehorende frequentie f , dan kun je het gemiddelde berekenen met de formule:

$$\bar{h} = \frac{\sum f \cdot s}{n}, \text{ waarbij } n = \sum f.$$

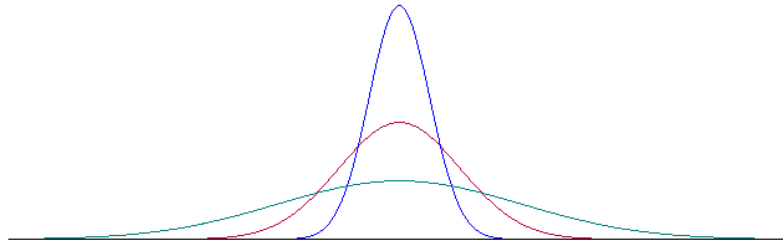
Dit is vooral handig als veel scores met hoge frequenties voorkomen.

- 46 a In de frequentietabel hiernaast staan de gegevens over de wiskundecijfers in klas 3 van het Amalia College. Vul op het werkblad de lege kolommen in en bereken daarmee voor de beide groepen het gemiddelde van *cijfwis*.
 b Vind jij dat de gemiddeldes bij de B- en de A/C-groep veel verschillen?

cijfwis	wiskundegroep			
	A/C		B	
	f	$f \times s$	f	$f \times s$
5	4		0	
6	12		9	
7	15		36	
8	11		44	
9	0		21	
10	1		1	
totaal	43		111	

Naast een centrummaat is ook de **spreiding** van belang.

In de figuur hieronder zie je drie frequentieverdelingen met hetzelfde gemiddelde en verschillende spreidingen. Het betreft bijvoorbeeld drie even grote bevolkingsgroepen met hetzelfde gemiddelde jaarinkomen, maar met sterk verschillende verdelingen. In die drie groepen zijn de spreidingen van de jaarinkomens zeer verschillend.



- 47 a Hoe zie je in de figuur dat de bevolkingsgroepen even groot zijn?
b Hoe zie je in de figuur dat het gemiddelde inkomen in de drie groepen hetzelfde is?
c Hoe zie je in de figuur dat de spreiding van de inkomens in de drie groepen verschillend is?

We willen de grootte van de spreiding in een getal uitdrukken. Dat wil zeggen dat we een **spreidingsmaat** willen hebben. Daarvoor zijn er verschillende mogelijkheden.

In paragraaf 1.3 (bladzijde 19 en 20) werd de spreiding vastgelegd door de kwartielafstand en gevisualiseerd in een boxplot.

Je kunt de spreiding ook goed uitdrukken met behulp van de deviaties (van het gemiddelde). Als er veel kleine deviaties zijn en weinig grote, is de spreiding gering. Maar de spreiding is groot als er grote positieve en negatieve deviaties voorkomen.

Een goede spreidingsmaat is de **gemiddelde absolute deviatie (gad)**¹⁾. Bij het aantal uren huiswerk van het B-groepje (zie opgave 42) zijn de deviaties -4 , -1 en $+5$; dus zijn de

absolute deviaties 4, 1 en 5, en is de spreidingsmaat *gad* gelijk aan $\frac{4+1+5}{3} = 3,3$.

- 48 a Bereken de *gad* voor het A/C-groepje.
b Welke van de twee groepen heeft de grootste spreiding met de *gad* als maatstaf?
- 49 Waarom kun je niet de *gemiddelde deviatie* als spreidingsmaat kiezen (dus zonder de absolute waarde te nemen)?

Een andere maat voor spreiding, de **variantie**, gebruikt de *kwadraten* van de deviaties. Door de kwadraten te nemen, heb je ook alleen met positieve getallen of nul te maken. Dit lijkt nodeloos ingewikkeld, maar het blijkt dat allerlei formules veel beter uitkomen als je met de kwadraten van de deviaties werkt in plaats van met de absolute waarden.

¹⁾ Het woord *absoluut* betekent hier dat de waarden positief (behalve 0) worden gemaakt. Een deviatie van 5 blijft onveranderd, maar een deviatie van -4 wordt 4. Je kunt ook zeggen dat de absolute deviatie let op de *afstand* van het gemiddelde en niet erop let of de afwijking naar links of naar rechts is.

<i>h</i> (aantal uur huiswerk)	$\bar{h}$ (gemiddelde)	<i>d</i> (deviatie)	d^2 (kwadraat)
10	11	-1	1
7	11	-4	16
16	11	5	25

De variantie (afgekort **var**) is het gemiddelde van de d^2 -waarden in de laatste kolom, dus

$$\frac{1+16+25}{3} = \frac{42}{3} = 14.$$

Voor n scores luidt de formule voor de variantie: $\text{var} = \frac{\sum d^2}{n}$.

- 50 a Bereken de variantie van de huiswerkuren in het A/C-groepje.
b Welke van de twee groepen heeft de grootste spreiding met de variantie als maatstaf?
- 51 Druk de huiswerktijd in plaats van in uren uit in kwartieren.
a Hoe groot worden dan de drie deviaties in het B-groepje?
b Bereken de variantie van de scores in kwartieren.
c Hoeveel keer zo groot is dit (vergeleken met de variantie van de scores in uren)?

Als je het huiswerk in uren rekent, is het gemiddelde ook in uren. De variantie is echter het gemiddelde van een kwadraat en daarmee zou die uur^2 als eenheid krijgen. Maar het ligt voor de hand om de spreiding van de huiswerkuren ook in uren te berekenen. Dat kan door de wortel van de variantie te nemen. Dat blijkt voor het latere formulewerk ook handig te zijn. De wortel uit de variantie is de **standaarddeviatie** of **standaardafwijking**, aangeduid met sd of met de enkele Griekse letter σ . In formulevorm:

$$sd = \sigma = \sqrt{\frac{\sum d^2}{n}}$$

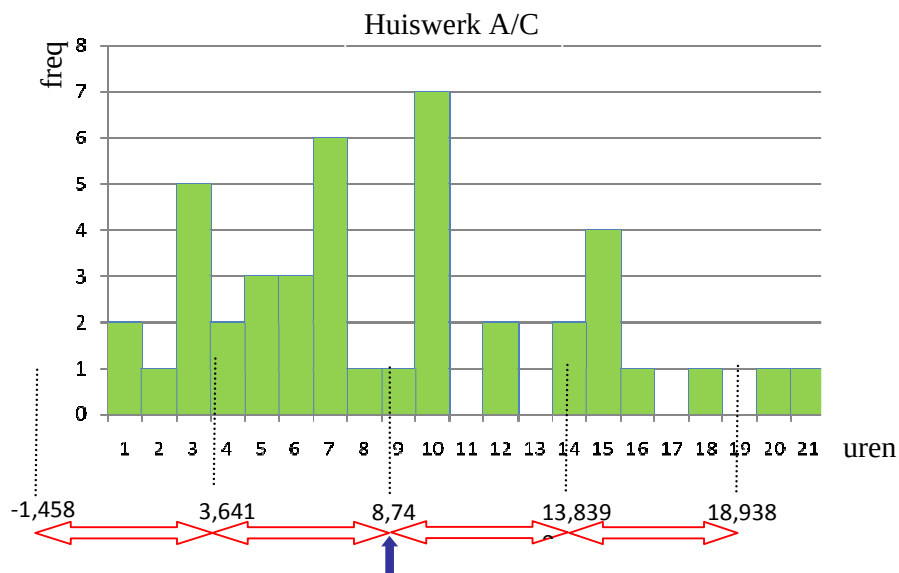
Voor het groepje met wiskunde B is de standaardafwijking van de huiswerktijd in uren $\sqrt{14} \approx 3,74$ uur.

- 52 Bereken de sd van de huiswerktijd in uren van de vier A/C-leerlingen.

In de praktijk bereken je de standaardafwijking (evenals het gemiddelde) met de Grafische Rekenmachine of met een computerprogramma. Het (laten) berekenen van gemiddelde en standaardafwijking is dus niet echt een probleem. Belangrijker is dat je een juiste voorstelling hebt van het begrip spreiding.

Bekijk daarom eens het histogram van de huiswerkinspanningen van de hele A/C-groep, op de volgende bladzijde.

Het gemiddelde 8,74 is met een pijltje aangegeven en de standaarddeviatie 5,099 is met pijlen langs de urenschaal afgepast, vanaf het gemiddelde. Eén pijl is één standaardafwijking naar links of naar rechts, twee pijlen zijn twee standaardafwijkingen naar links of naar rechts. In dit geval reiken de pijlen tot $8,74 - 2 \times 5,099 = -1,458$, tot $8,74 - 5,099 = 3,641$, tot $8,74 + 5,099 = 13,839$ en tot $8,74 + 2 \times 5,099 \approx 18,938$.



Voor veel verdelingen gelden de volgende vuistregels:

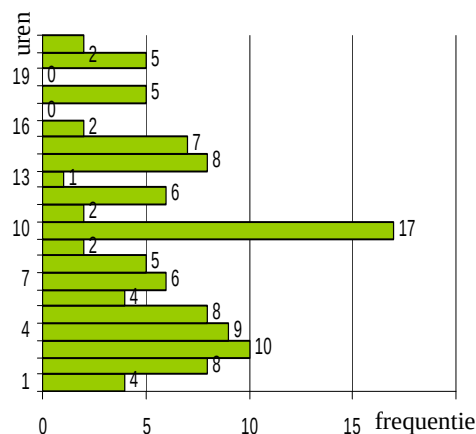
- Tussen het gemiddelde-min-sd en het gemiddelde-plus-sd ligt ongeveer 68% van de gehele verdeling.
- Tussen het gemiddelde-min-2-keer-sd en het gemiddelde-plus-2-keer-sd ligt ongeveer 95% van de gehele verdeling.

Hoe nauwkeurig die 68% en die 95% kloppen hangt erg af van de verdeling zelf. Komen in het midden (dicht bij het gemiddelde) veel scores voor en neemt hun aantal af naar de uiteinden, dan klopt deze percentages heel aardig.

We controleren de eerste vuistregel voor de huiswerkuren van de A/C-groep. Daarvoor moeten we bepalen hoeveel A/C-leerlingen meer dan 3,58 uur en minder dan 13,90 uur aan huiswerk besteden. Dat zijn bijna alle leerlingen die 4 uur huiswerk maken, alle leerlingen die 5, 6, 7, 8, 9, 10, 11, 12 of 13 uur besteden en nog een deel van de leerlingen die 14 uur aan huiswerk besteden. Hoeveel precies is niet onmiddellijk duidelijk. Voor de berekening gaan we als volgt te werk: $0,92 \cdot 2 + 3 + 3 + 6 + 1 + 1 + 7 + 0 + 2 + 0,40 \cdot 2 = 25,64$.

- 53 a Ga na wat het idee achter deze berekening is.
 b Klopt de eerste vuistregel ongeveer?
 c Bereken hoeveel procent van de aantallen uren huiswerk in de A/C-groep ligt tussen *gemiddelde* - $2 \times sd$ en *gemiddelde* + $2 \times sd$.
 d Klopt de tweede vuistregel ongeveer?

- 54 Hiernaast staat een 'horizontaal' histogram van het aantal uren huiswerk in de B-groep.
 a Bereken aan de hand daarvan het gemiddelde en de standaarddeviatie.
 b Bereken het percentage tussen *gemiddelde* - *sd* en Bereken ook het percentage tussen *gemiddelde* - $2sd$ en *gemiddelde* + $2sd$.
 Tip: Voer de aantallen uren en de frequenties in op de GR in gemiddelde en de sd berekenen.



Voorbeeld: het IQ

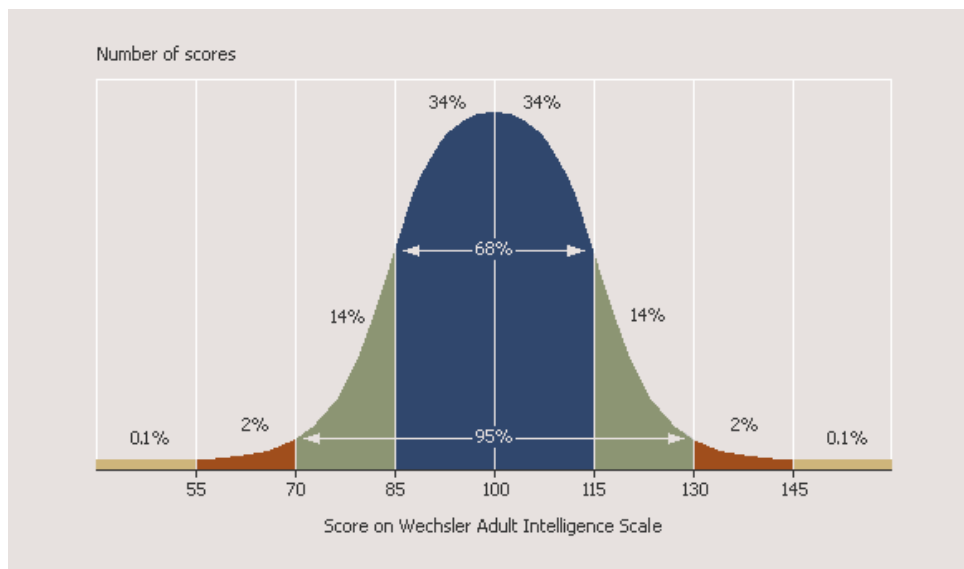
Oorspronkelijk definieerde Henry Goddard het intelligentiequotiënt als volgt:
$$IQ = \frac{\text{verstandelijke leeftijd}}{\text{werkelijke leeftijd}} \times 100.$$

Voor kinderen was dit aanvankelijk een bruikbare definitie, maar voor volwassenen niet. David Wechsler ontwikkelde een IQ-meting door de prestatie van de proefpersoon te vergelijken met de scores van grote normgroepen, een principe dat sindsdien algemeen verspreid werd.

Intelligentietests worden zó ontworpen dat de verdeling van IQ-scores ongeveer 'normaal' is, met als gemiddelde 100 en als standaardafwijking 15. Normaal betekent "klokvormig" zoals in de figuur hieronder: met de meeste waarnemingen in het midden van de verdeling, en naar links en rechts duidelijk afnemende aantallen. Het gemeten IQ wordt gezien als een *schatting* van het ware, maar onbekende, IQ. Herhaalde proeven bij een zelfde persoon wijzen uit dat diens testuitslag soms tot twintig punten kan verschillen met een eerdere test, door allerlei oorzaken zoals gezondheid, vermoeidheid, stress en gewenning aan materiaal en situatie.

De indeling van IQ-waarden is als volgt:

- 145-160 Zeer hoogbegaafd
- 130-144 Hoogbegaafd
- 116-130 Boven gemiddeld begaafd
- 101-115 Begaafd
- 100 Gemiddeld
- 85-99 Beneden Gemiddeld
- 70-84 Moeilijk lerend
- 55-69 Lichte verstandelijke beperking
- 40-54 Matige verstandelijke beperking
- 25-39 Ernstige verstandelijke beperking
- < 25 Zeer ernstige verstandelijke beperking

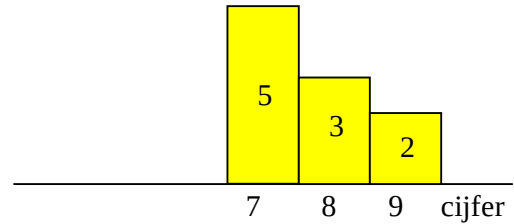


- 55 In het diagram zie je een verdeling van de IQ-scores van volwassenen.
- a Wat is de mediaan?
 - b Geef een schatting van het eerste en het derde kwartiel. (Dat hoeft niet zo precies.)
 - c De sd is 15. Hoeveel procent ligt tussen *gemiddelde*+sd en *gemiddelde*-sd, hoeveel tussen *gemiddelde*+2sd en *gemiddelde*-2sd en hoeveel tussen *gemiddelde*+3sd en *gemiddelde*-3sd?

Begrepen?

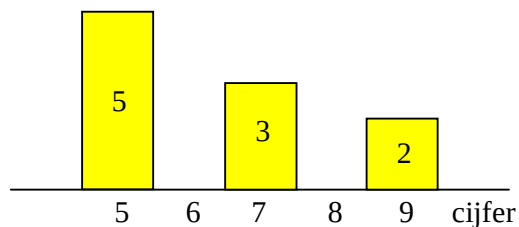
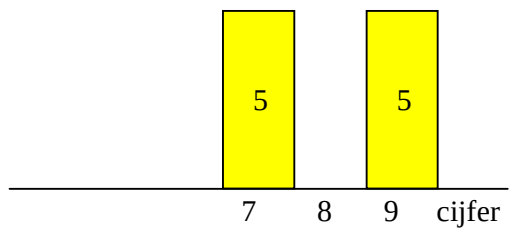
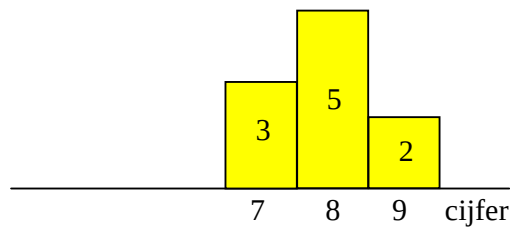
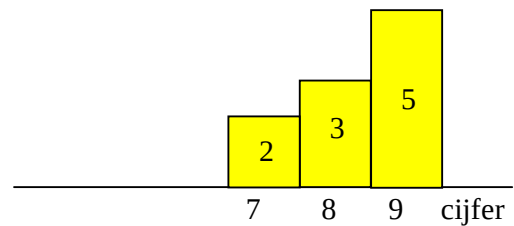
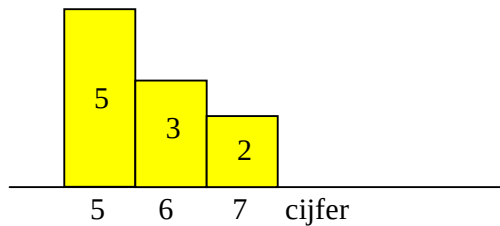
In een klas van 10 leerlingen zijn de proefwerkcijfers als volgt. Het gemiddelde is 7,7 en de sd is 0,78.

- Kun je dit narekenen?



Hieronder staan nog vijf mogelijke resultaten van het proefwerk. Het is niet de bedoeling sd's van elk van deze apart uit te rekenen, maar om die te vergelijken met de gegeven sd van 0,78 van het voorbeeld hierboven.

- Zeg van elk van deze vijf of de sd groter is dan, kleiner is dan of gelijk is aan de sd van het voorbeeld hierboven.



1.8 Effectgrootte

Is er verschil?

Heeft het gebruik van kunstmest bij appelbomen zin? Om daar achter te komen worden twee groepen appelbomen bekeken: in groep A wordt kunstmest gebruikt en in groep B niet. Het blijkt dat de appels in groep A gemiddeld een diameter van 6,8 cm hebben met een standaardafwijking van 0,28 cm. De appels in groep B hebben een diameter van 6,2 cm met een standaardafwijking van 0,20 cm.

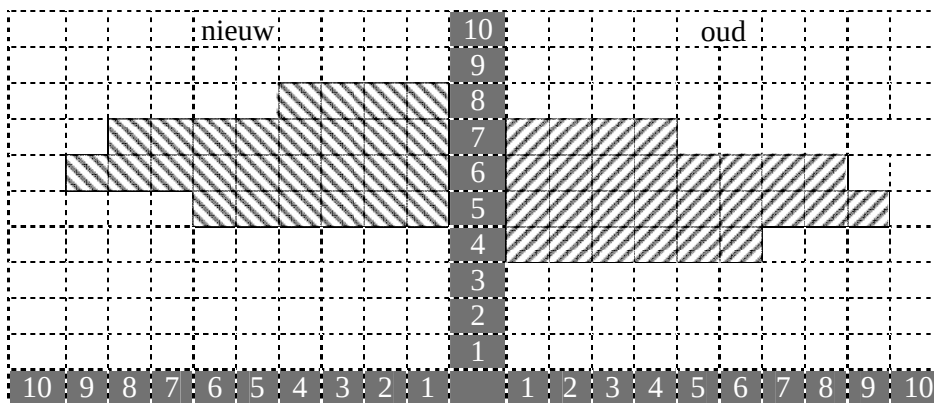
Kan op grond van dit resultaat beslist worden dat het gebruik van kunstmest zinvol is, of is het verschil daarvoor te gering? (Kleine verschillen kunnen altijd wel optreden door toevallige verschillen in omstandigheden.)

Dit is een vraag waar de statistiek antwoord op wil geven. We willen het *effect* van het gebruik van kunstmest in een getal uitdrukken. Als dat getal groot is, zullen we concluderen dat kunstmest grotere appels geeft, als het getal klein is, zullen we dat niet zeggen (en natuurlijk blijven er twijfelgevallen).

- 56 a Stel dat groep A als gemiddelde 7,0 cm met standaardafwijking 0,28 zou geven, was dan het effect van het gebruik van kunstmest groter of kleiner dan bij het bovenbeschreven resultaat?
- b Stel dat groep A als gemiddelde 6,8 cm met standaardafwijking 0,23 zou geven, was dan het effect van het gebruik van kunstmest groter of kleiner dan bij bovenbeschreven resultaat?

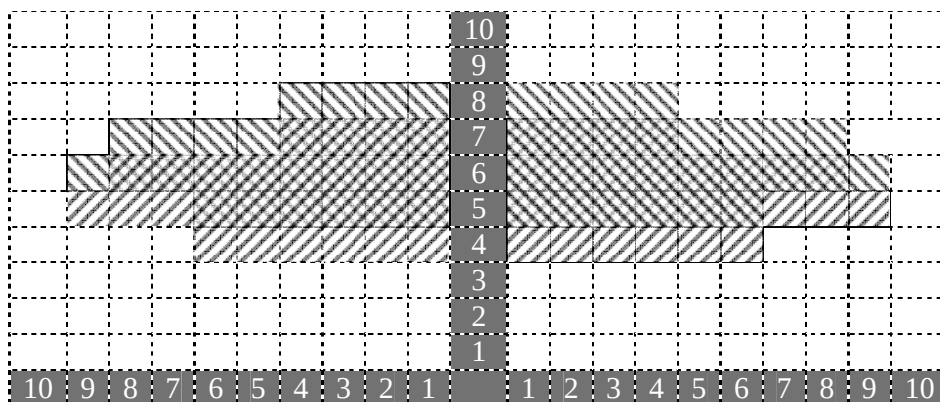
overlap

Bij een onderwijskundig onderzoek wordt in twee vergelijkbare klassen een wiskundig begrip op twee verschillende manieren uitgelegd: een traditionele manier en een nieuwe manier. In de twee klassen werd de uitleg afgesloten met dezelfde toets.



Het dubbele histogram laat de (denkbeeldige) toetsresultaten van de twee klassen zien. De klas waarvan links het histogram staat had de nieuwe manier van uitleg gekregen. De cijfers in het rechter histogram zijn precies 1 punt lager dan in het linker. Het gemiddelde van de rechter klas is dus precies 1 punt lager dan dat van de linker klas. De standaarddeviaties zijn gelijk (Kun je uitleggen waarom?); in dit geval is $sd = 0,99$. De nieuwe manier van uitleg lijkt een duidelijk effect te hebben. Hoe groot het effect is, gaan we uitdrukken in een getal.

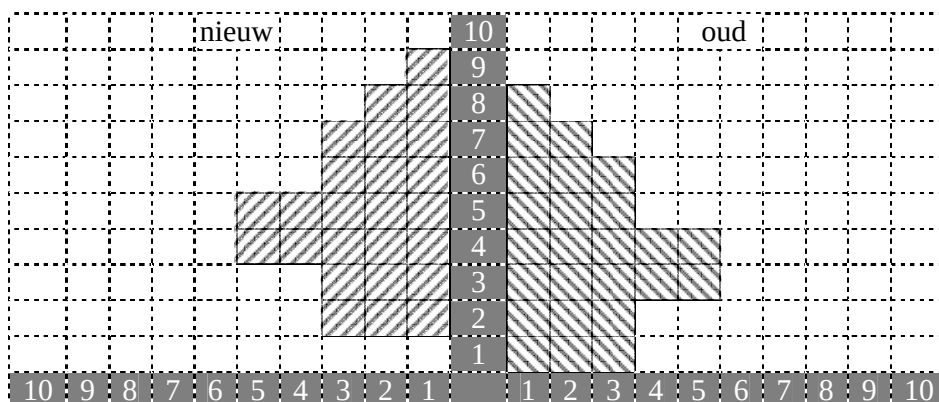
Daartoe leggen we de twee histogrammen over elkaar en letten op de overlap.



Er zitten 18 van de 27 leerlingen in de overlap, dat is 67%; dus 33% zit niet in de overlap.

Bij 100% overlap zou het effect van de nieuwe uitleg nihil zijn. Bij 0% overlap zou het effect van de nieuwe uitleg zeer groot zijn. Het percentage dat niet in de overlap zit wordt genoteerd met de Griekse letter Δ . In dit geval is $\Delta = 0,33$ (33%).

- 57 Het experiment wordt herhaald in twee vergelijkbare klassen op een andere school. Hieronder staat het (weer denkbeeldige) resultaat.



- Hoeveel verschillen de gemiddeldes tussen de twee klassen? En de standaardafwijkingen?
- Waarom zie je dat op de tweede school de standaardafwijking van de toetsresultaten groter is dan op de eerste school?

De standaardafwijking is nu 1,93.

- Leg op het werkblad de histogrammen over elkaar.
- Hoeveel procent van de leerlingen zit in het gemeenschappelijke deel?
- Hoe groot is Δ ?
- Op welke school heeft de nieuwe uitleg het meeste effect gehad?

Het percentage Δ dat *niet* in de overlap zit is een maat voor het effect dat een nieuwe uitleg heeft. Als $\Delta = 0$, is er geen effect. Hoe groter Δ , des te groter het effect. De groepen die je vergelijkt moeten even groot zijn.

effectgrootte

Het effect van de nieuwe uitleg kan ook op een andere manier worden gemeten. Daarbij is niet alleen het verschil van de gemiddelde toetscijfers van belang, maar ook de standaardafwijkingen in beide groepen.

J. Cohen stelde in 1962 voor de effectgrootte bij een vergelijking van twee groepen als volgt te berekenen: neem het verschil tussen de gemiddeldes en deel dat door de standaardafwijking (als tenminste de standaardafwijkingen in beide groepen gelijk zijn; neem anders het gemiddelde van de standaardafwijkingen). Hij gebruikte de letter D voor deze effectgrootte.

$$\text{Effectgrootte} = D = \frac{\text{verschil tussen de gemiddeldes}}{\text{gemiddelde van de standaardafwijkingen}}$$

Hierbij is er altijd sprake van een groep waarop iets nieuws wordt uitgetoetst en een groep waarop dat niet gebeurt (de *controlegroep*). De letter D is van “differentie”, wat “verschil” betekent.

Waardoor in deze formule gedeeld moet worden is onderwerp van discussie tussen statistici. Volgens sommigen kan beter gedeeld worden door de standaardafwijking van de controlegroep. Wij zullen bovenstaande definitie hanteren.

- 58 a Ga met een berekening na dat de effectgrootte in de eerste school (opgave 55) ongeveer 1 is.
b Bereken de effectgrootte in de tweede school (opgave 56).
c In welke school had de nieuwe uitleg het meeste effect?

Als waardering van D is vrij gangbaar:

- $D \leq 0,4$ gering
- $0,4 < D \leq 0,8$ middelmatig
- $0,8 < D \leq 1,5$ groot
- $D > 1,5$ erg groot

- 59 a Bereken de effectgrootte van kunstmest in de situatie van het begin van deze paragraaf.
b Is het effect van kunstmest groot?
- 60 Waarom wordt er in de formule voor de effectgrootte überhaupt gedeeld (door het gemiddelde van de standaardafwijkingen)? Waarom is het verschil tussen de gemiddeldes niet goed? In deze opgave wordt dat uitgelegd.
Stel dat de cijfers niet op een schaal van 1 t/m 10 zouden zijn gegeven, maar op een schaal van 1 t/m 100. We kijken naar de eerste school.
a Wat zou dan het verschil tussen de gemiddeldes zijn?
b Wat zou de standaardafwijking dan zijn?
c Wat is dan dus de effectgrootte?

Dankzij het feit dat gedeeld wordt door de standaardafwijking, doet de “schaal” waarop de cijfers worden gegeven er niet toe. Als niet gedeeld zou worden door de standaardafwijking zou je niet kunnen zeggen of een effectgrootte van bijvoorbeeld 1,9 groot is of niet.

- 61 Op het Amalia College telt in vwo4 de A/C-groep 40 leerlingen en de B-groep 30 leerlingen. In de volgende tabel zijn van vijf variabelen voor beide groepen het gemiddelde en de standaardafwijking gegeven.

Variabele		A/C (n = 40)	B (n = 30)	Effectgrootte D
leeftijd	gem.	15,8	15,1	
	sd	0,9	0,7	
cijfgem	gem.	7,25	7,43	
	sd	0,675	0,772	
cijfwis	gem.	6,25	7,67	
	sd	0,968	1,164	
kunstbel	gem.	3,6	2,7	
	sd	1,02	1,13	
huiswerk	gem.	8,57	10,47	
	sd	3,41	3,15	

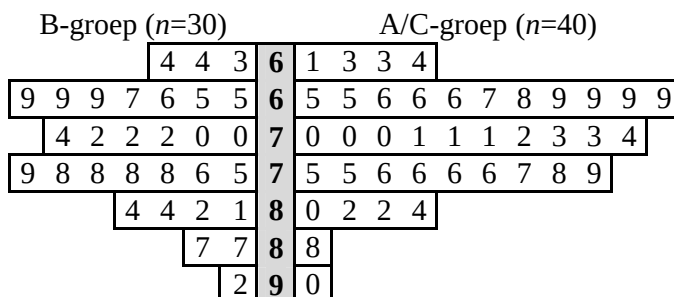
- Bereken voor elke variabele de effectgrootte D van de groep. Zie werkblad.
- Op welke variabele verschillen de twee groepen het meest?
- Op welke variabele(n) verschillen de twee groepen maar weinig, vind jij?

- 62 Hieronder zie je een dubbel steel-blad-diagram van de scores op *cijfgem* in de wiskunde-A/C- en de wiskunde-B-groep van het Amalia College.



- Bereken de effectgrootte D van de groep op de variabele *cijfgem*.

Hieronder zijn hokken om de bladeren van het steelbladdiagram getekend. Dan gaat het op twee histogrammen lijken.



- Wat zijn de klassengrenzen bij deze histogrammen?
- Waarom kun je in dit geval niet de verschilmaat Δ met de gegevens van deze histogrammen berekenen?

- 63** We illustreren de effectgrootte aan de hand van de lengte van jongens en meisjes. Zet je een groep van 18-jarige meisjes naast een groep 18-jarige jongens, dan zie je het lengteverschil onmiddellijk. Op 14-jarige leeftijd is het verschil nog nauwelijks waarneembaar. Tot de leeftijd van 13 jaar zijn jongens en meisjes in Nederland vrijwel even lang, maar daarna ontwikkelt zich een verschil. Meisjes van 14 jaar zijn gemiddeld 165 cm lang en jongens van deze leeftijd 168 cm. De standaardafwijking zowel bij de meisjes als bij de jongens is op die leeftijd 7 cm.
- Bereken de effectgrootte van de sekse op de lengte bij 14-jarigen.
 - Is hier sprake van een klein, matig, groot of erg groot verschil?

Op 18-jarige leeftijd zijn de gemiddeldes 170 cm en 183 cm, terwijl de standaardafwijkingen bij beide groepen 7 cm blijft.

- Wat is nu de effectgrootte en hoe zou je de grootte van dat verschil nu kwalificeren?

-  **64** Voor de jongens en meisjes in 4 vwo van het Amalia College zijn de gemiddelde lengtes: 169,9 en 176,5 cm. De bijbehorende standaardafwijkingen zijn 8,6 en 8,9 cm.
- Is hier sprake van een klein, matig, groot of erg groot verschil?

-  **65** Iemand beweert: “Bij de keuze in klas 3 voor wel of geen wiskunde B speelt op het Amalia College het cijfer voor wiskunde een grote rol”. Zie opgave 61 voor de gegevens.
- Wat vind je van die uitspraak?

-  **66**
 - Volgens sommige docenten zijn er duidelijke verschillen in het maken van huiswerk tussen jongens en meisjes. Onderzoek hoe dat zit in het bestand [school.vus](http://school.vus.nl). Kies *huiswerk* en splits naar *geslacht*. Berekeningen zoals de sd kun je uitvoeren in Dataplot onder de knop . Bereken zelf de effectgrootte.
 - Ga ook na hoe het zit met verschillen tussen de profielen ten aanzien van de tijd die besteed wordt aan huiswerk.
 - Het lijkt vreemd om te letten op lengteverschillen tussen de profielen. Toch is in één profiel de gemiddelde lengte duidelijk hoger. Hoe is dat te verklaren?

Begrepen?

- Een proefwerk was erg slecht gemaakt. Daarom gaf de lerares extra uitleg met daarna een herkansing. Bij de herkansing scoorde iedereen 2 punten hoger dan bij het eerste proefwerk.
Wat moet je nog weten om de effectgrootte D van de extra uitleg op het proefwerk-resultaat uit te kunnen rekenen?
Wat is de grootste mogelijke waarde van de effectgrootte D ?
- Van twee even grote groepen zijn de gemiddeldes van een variabele bekend. Hoe groter Δ , des te de overlap, des te is de spreiding, des te is D .

1.9 Onderzoek 2



In de volgende opdrachten ga je met het programma VUStat het bestand school.vus onderzoeken. Dit bestand bevat de gegevens van twee scholen en geeft daarmee een beeld van het verschil tussen wiskunde-A/C-leerlingen en wiskunde-B-leerlingen.

VUStat heeft twee geschikte modules voor statistisch onderzoek: Statistiek en Dataplot. In de module Statistiek werk je met zogenaamde ruwe data die in een datatabel/datamatrix staan. In de module Dataplot werk je met geordende data in de vorm van frequentietabellen. In beide modules kun je berekeningen en afbeeldingen maken. Je kunt tussen deze modules switchen. Een frequentietabel die je in Statistiek maakt kun je met een druk op de knop naar Dataplot sturen en daar desgewenst opslaan of er verder mee werken.

Hoe maak je berekeningen in VUStat?

Net als in paragraaf 1.6 gaan we aan de hand van het voorbeeld *kunstbel*, gesplitst in A/C en B, het maximale percentageverschil berekenen.

Open in de module Statistiek het bestand *school.vus* en volg onderstaande stappen.

1. Kies in het menu Tabel > Frequentietabel en selecteer de variabele *kunstbel*.
2. Klik op de knop Splitsen om te groeperen op de variabele *wisgroep*.
3. Vink Procenten en Cumulatief aan.
4. Klik op de knop Dataplot om de tabel naar Dataplot te sturen.
5. Klik in de eerste lege kolom van Dataplot en maak de variabele *deltap* aan, type geheel getal.
6. Klik rechts in de kolom *deltap* en klik op Omrekenen deltap.
7. Vul door aanklikken de formule $\text{abs}([A/C_cum\%] - [B_cum\%])$ in. ¹⁾

Je kunt nu het maximale cumulatieve percentageverschil *max.Vcp* vaststellen.

Zowel in Statistiek als in Dataplot kun je het gemiddelde en de standaardafwijking berekenen. In Statistiek kies je in het menu Meer Statistiek > Centrummaten. Je kunt ook een variabele splitsen in deelgroepen en daarvan de centrummaten aflezen.

In Dataplot is een knop  of je kiest Opties > Centrummaten. Denk er om dat je met frequenties rekening moet houden.

- 67** Presteerden A/C-leerlingen in klas 3 even goed in wiskunde als B-leerlingen?
Bereken - om deze vraag te beantwoorden - het maximale cumulatieve percentageverschil van *cijfwis* voor de A/C - en de B-groep. Hoe interpreteer je het gevonden verschil?
- 68** Is wiskunde-A/C 'meisjeswiskunde' en wiskunde-B 'jongenswiskunde'?
Bepaal hiervoor de odds-ratio voor man : vrouw bij de A/C- en de B-groep.
Hoe interpreteer je het gevonden resultaat?
Laat een tabel maken van *geslacht* gesplitst op *wisgroep*.
- 69** Om de vraag van opgave 66 te kunnen beantwoorden kun je ook kijken naar de effectgrootte *D*.
a Bereken *D* met behulp van VUStat. Bereken daarvoor de gemiddeldes en de standaarddeviaties uitrekenen van *cijfwis* voor de beide groepen.
Hoe interpreteer je het gevonden resultaat? Vergelijk dit met je antwoord bij opgave 66.
b Bereken ook het percentage Δ dat niet in de overlap zit. Omdat de groepen niet even groot zijn, moet je met percentages werken.

¹⁾ De wiskunde functie *abs* neemt de absolute waarde; deze laat positieve getallen en nul onveranderd en maakt van negatieve getallen hun tegengestelde.

1.10 Samenvatting van paragraaf 1.6 t/m 1.9

We vergelijken twee groepen wat een bepaalde variabele betreft.

Het **maximale cumulatieve percentageverschil** $\max.V_{cp}$ tussen twee groepen bereken je als volgt:

- Bepaal apart voor de groepen de cumulatieve frequenties;
- Zet apart voor de groepen de cumulatieve frequenties om in cumulatieve percentages;
- Bepaal per waarde van de variabele het (absolute) verschil V_{cp} van de cumulatieve percentages van de twee groepen.
- $\max.V_{cp}$ is van al die verschillen het grootste.

Als waardering van $\max.V_{cp}$ is gangbaar:

- als $\max.V_{cp} \leq 15\%$, is het verschil gering,
als $15\% < \max.V_{cp} \leq 30\%$, is het verschil middelmatig,
als $\max.V_{cp} > 30\%$, is het verschil groot.
-

We vergelijken twee groepen (bijvoorbeeld A/C en B) wat een bepaalde variabele betreft. De variabele heeft twee waarden (bijvoorbeeld m en v).

De odds van m tegen v in de B-groep is $20 : 10 = 2$

De odds van m tegen v in de A/C-groep is $18 : 22 \approx 0,82$

De **odds-ratio** of **ongelijkheidsverhouding** van m tegen v in de twee groepen is $2 : 0,82 \approx 2,2$

Als dit quotiënt kleiner dan 1 is, nemen we het omgekeerde.

geslacht	A/C	B	tot
m	18	20	38
v	22	10	32
totaal	40	30	70

Als waardering van de odds-ratio is gangbaar:

- als $\text{odds-ratio} \leq 2$, is het verschil gering,
als $2 < \text{odds-ratio} \leq 3$, is het verschil middelmatig,
als $\text{odds-ratio} > 3$, is het verschil groot.
-

Het **gemiddelde** van een verdeling is op te vatten als de ‘evenwichtswaarde’. Dat wil zeggen dat de som van alle afwijkingen ten opzichte van die waarde 0 is.

In formule:

$$\bar{x} = \frac{\sum x}{n}, \text{ waarin } x \text{ de scores zijn en } n \text{ het totaal aantal scores is,}$$

of

$$\bar{x} = \frac{\sum x \cdot f}{n}, \text{ waarin } x \text{ de verschillende scores zijn met bijbehorende frequenties } f \text{ en } n \text{ de som van de frequenties is.}$$

De **standaardafwijking** of **standaarddeviatie** is een maat voor de spreiding van de scores rond het gemiddelde. In formule:

$$sd = \sigma = \sqrt{\frac{\sum d^2}{n}}$$

waarin d de deviaties (= afwijkingen) van het gemiddelde zijn; $d = x - \bar{x}$.

Voor veel verdelingen gelden de volgende *vuistregels*:

- tussen $\bar{x} - \sigma$ en $\bar{x} + \sigma$ ligt ongeveer 68% van alle scores,
- tussen $\bar{x} - 2\sigma$ en $\bar{x} + 2\sigma$ ligt ongeveer 95% van alle scores.

We vergelijken twee even grote groepen wat een variabele betreft. Van de variabelen is de verdeling bij beide groepen bekend.

Δ is een maat voor het verschil tussen de groepen aan de hand van de overlap.

Δ wordt als volgt bepaald:

- leg de verdelingen over elkaar,
- bepaal de overlap,
- bepaal hoeveel procent van een groep niet in de overlap zit,
- dat percentage is Δ .

We vergelijken twee groepen wat een variabele betreft. Van de variabelen is de verdeling bij beide groepen bekend. De groepen hoeven niet even groot te zijn.

De **effectgrootte** is een maat voor het verschil tussen de groepen aan de hand van de gemiddeldes en de standaarddeviaties van de verdelingen.

In formule:

$$D = \frac{\text{afstand tussen de gemiddeldes}}{\text{gemiddelde van de standaardafwijkingen}} \cdot$$

Als waardering van D is vrij gangbaar:

- als $D \leq 0,4$, is het verschil gering,
- als $0,4 < D \leq 0,8$, is het verschil (middel)matig,
- als $0,8 < D \leq 1,5$, is het verschil groot,
- als $D > 1,5$, is het verschil erg groot.