4. Statistische uitspraken doen

Boekje 4 havo wiskunde A, domein E: Statistiek

Verantwoording

Dit lesmateriaal is ontwikkeld in het kader van de nieuwe examenprogramma's zoals voorgesteld door de commissie Toekomst Wiskunde Onderwijs (cTWO) en herzien door SLO.

De lessenserie Statistiek op een groot gegevensbestand en de diagnostische computertoets zijn mede mogelijk gemaakt door het Centraal Bureau voor de Statistiek. De auteurs bedanken het CBS en in het bijzonder Lieke Stroucken van de afdeling 'CBS in de klas' voor de samenwerking.

Auteurs: Erik van Barneveld, Wouter Boer, Carel van de Giessen, Peter Kop, Heleen van der Ree, Henk Reuling, Frits Spijkers, Tanja Stroosma, Anneke Verschut

Met medewerking van: Nico Alink, Martine de Klein (eindredactie)

Informatie: SLO
Afdeling: tweede fase
Postbus 2041, 7500 CA Enschede
Telefoon (053) 4840 661
Internet: www.slo.nl
E-mail: tweedefase@slo.nl

Overzicht lesmateriaal in het domein Statistiek

1. Kijken naar data

§ 1.1 Wat is statistiek?
§ 1.2 Data
§ 1.3 Diagrammen
§ 1.4 Interpretaties
§ 1.5 Overzicht

2. Data en datasets verwerken

§ 2.0 Begrippenlijst
§ 2.1 Data presenteren
§ 2.2 Verbanden tussen datarepresentaties
§ 2.3 Frequentieverdelingen typeren
§ 2.4 Twee groepen vergelijken
§ 2.5 Samenhang tussen twee variabelen

3. Data verwerven

§ 3.0 Pas op voor valkuilen
§ 3.1 Onderzoeks- en enquêtevragen
§ 3.2 Steekproeven en fouten
§ 3.3 Standaardafwijking
§ 3.4 Steekproeffout: variatie bij steekproeven
§ 3.5 Normale verdeling
§ 3.6 Toevallige steekproeffouten in getallen
§ 3.7 Terugblik op boekje 3

4. Statistische uitspraken doen

§ 4.1 Voorkennis
§ 4.2 Doel van deze module
§ 4.3 Populatieproportie
§ 4.4 Populatiegemiddelde
§ 4.5 Verschil tussen twee groepen
§ 4.6 Samenhang tussen twee kwantitatieve variabelen
§ 4.7 Gemengde opgaven
§ 4.8 Terugblik
§ 4.9 Lessenserie: Statistiek op een groot gegevensbestand
§ 4.10 Diagnostische computertoets

Inhoud

Overzicht lesmateriaal in het domein Statistiek 3

§ 4.1 Voorkennis 6

§ 4.2 Doel van deze module 9

§ 4.3 Populatieproportie 10

§ 4.3.1 Introductie 10

§ 4.3.2 Centrale vraag 10

§4.3.3 Antwoord op de centrale vraag 11

§ 4.3.4 Oefenen 12

§ 4.3.5 Om te onthouden 13

§ 4.3.6 Geïntegreerd oefenen 14

§ 4.4 Populatiegemiddelde 15

§ 4.4.1 Introductie 15

§ 4.4.2 Centrale vraag 15

§ 4.4.3 Antwoord op de centrale vraag 16

§ 4.4.4 Oefenen 17

§ 4.4.5 Om te onthouden 18

§ 4.4.6 Geïntegreerd oefenen 18

§ 4.5 Verschil tussen twee groepen 21

§ 4.5.1 Op een nominale variabele (phi) 21

§ 4.5.1.1 Introductie 21

§ 4.5.1.2 Centrale vraag 21

§ 4.5.1.3 Antwoord op de centrale vraag 22

§ 4.5.1.4 Oefenen 24

§ 4.5.1.5. Om te onthouden 25

§ 4.5.1.6 Geïntegreerd oefenen 25

§ 4.5.2 Op een ordinale variabele ($\max V_{cp}$) 27

§ 4.5.2.1 Introductie 27

§ 4.5.2.2 Centrale vraag 27

§ 4.5.2.3 Antwoord op de centrale vraag 27

§ 4.5.2.4 Oefenen 30

§ 4.5.2.5 Om te onthouden 31

§ 4.5.3 Op een kwantitatieve variabele met effectgrootte 32

§ 4.5.3.1 Introductie 32

§ 4.5.3.2 Centrale vraag 32

§ 4.5.3.3 Antwoord op de centrale vraag 32
§ 4.5.3.4 Oefenen 33
§ 4.5.3.5 Om te onthouden 33
§ 4.5.4 Op een kwantitatieve variabele (vergelijken van boxplots) 34
§ 4.5.4.1 Introductie 34
§ 4.5.4.2 Centrale vraag 34
§ 4.5.4.3 Antwoord op de centrale vraag 35
§ 4.5.4.4 Oefenen 37
§ 4.5.4.5 Om te onthouden 38
§ 4.5.4.6 Geïntegreerd oefenen 38
§ 4.6 Samenhang tussen twee kwantitatieve variabelen 43
§ 4.6.1 Correlatiecoëfficiënt 43
§ 4.6.1.1 Introductie 43
§ 4.6.1.2 Centrale vraag 44
§ 4.6.1.3 Antwoord op de centrale vraag 44
§ 4.6.1.4 Oefenen 48
§ 4.6.1.5 Om te onthouden 49
§ 4.6.2 Trendlijn 49
§ 4.6.2.1 Introductie 49
§ 4.6.2.2 Centrale vraag 50
§ 4.6.2.3 Antwoord op de centrale vraag 50
§ 4.6.2.4 Oefenen 51
§ 4.6.2.5 Om te onthouden 53
§ 4.7 Gemengde opdrachten 54
§ 4.8 Terugblik 58
§ 4.9 Lessenserie: Statistiek op een groot gegevensbestand 59
§ 4.10 Diagnostische computertoets 68

§ 4.1 Voorkennis

Om deze module te begrijpen, moet je weten wat de volgende begrippen betekenen:

- Statistische cyclus.
- Populatie en steekproef.
- Variabelen en hun meetniveau.

Deze begrippen komen hieronder (nogmaals) aan bod.

Statistische cyclus

- We beginnen een statistisch onderzoek met een vraag die (alleen) met statistische gegevens kan worden beantwoord.
- Dan stellen we vast over welke populatie (doelgroep) het onderzoek gaat. De leden van de populatie zijn de elementen waar het onderzoek betrekking op heeft.
- Vervolgens stellen we nauwkeurig vast op welke statistische variabele het onderzoek betrekking heeft.
- Daarna verzamelen we de bij die variabele passende data (gegevens). Dit kunnen gegevens zijn van de gehele populatie of van een deel van de populatie (steekproef).
- Vervolgens ordenen en analyseren we de verzamelde gegevens om meer overzicht te krijgen.
- Tenslotte proberen we een conclusie te trekken.

Populatie en steekproef

- In het algemeen willen we iets weten over alle elementen waarop het onderzoek betrekking heeft. Vaak zijn de elementen personen, maar het kunnen ook auto's zijn of lampen etc.
  Soms is het mogelijk om al die elementen (de gehele populatie) te onderzoeken, maar vaak is dat onmogelijk, omdat het te veel tijd zou kosten en/of te duur zou zijn. Als dat het geval is, beperken we ons tot een steekproef uit de populatie waarin we zijn geïnteresseerd.
- De steekproefomvang is het aantal elementen in de steekproef.
- De steekproefopzet is de manier waarop we bepalen welke elementen uit de populatie in de steekproef terechtkomen.
- Een steekproef is aselect als deze geen 'voorkeur' heeft voor bepaalde soorten elementen.
  Met andere woorden: elk element uit de populatie heeft dezelfde kans heeft om in de steekproef te zitten.
- We noemen een steekproef representatief voor de populatie waaruit deze getrokken is wanneer de steekproef een vergelijkbare verdeling vertoont op alle relevante variabelen.
  De beste manier om representativiteit te garanderen is het trekken van een aselecte steekproef.

Variabelen en hun meetniveau

- Van belang in deze module is het onderscheid tussen kwalitatieve en kwantitatieve variabelen.
- Kwalitatieve variabelen zijn variabelen waarvan we de uitkomsten niet uitdrukken in getallen (met een getalsmatige betekenis).
  Bij kwantitatieve variabelen drukken we de uitkomsten wel uit in getallen (met getalsmatige betekenis).
- Binnen de kwalitatieve variabelen maken we een onderscheid in nominale en ordinale variabelen.
- Bij een variabele op nominaal niveau vormen de mogelijke waarden niet meer dan labels voor de categorieën. Een voorbeeld van een variabele op nominaal niveau is geslacht. Deze variabele heeft twee mogelijke waarden: man of vrouw. Man en vrouw kun je zien als labels voor de twee verschillende geslachten.
- Bij een variabele die op ordinaal niveau is gemeten heeft de ordening van de waarden van laag naar hoog een bepaalde betekenis. Een voorbeeld van een variabele op ordinaal niveau is bijvoorbeeld werkhouding als dit wordt beoordeeld via de mogelijke waarden onvoldoende, matig, voldoende en goed.

Bij het eindexamen zal een formuleblad afgedrukt worden.
Deze staat hieronder afgedrukt en leer je ook te gebruiken.

Formuleblad bij Centraal Schriftelijke Eindexamen

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 $\overline{X} \pm 2 \cdot \frac{S}{\sqrt{n}}$

met $\overline{X}$ het steekproefgemiddelde, n de steekproefomvang en S de steekproefstandaardafwijking.

Vuistregels bij de grootte van het verschil van twee groepen

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

- $phi < -0{,}4$ of $phi > 0{,}4$: groot verschil.
- $-0{,}4 < phi < -0{,}2$ of $0{,}2 < phi < 0{,}4$: middelmatig verschil.
- $-0{,}2 < phi < 0{,}2$: gering verschil.

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

- $\max V_{cp} > 40$: groot verschil.
- $20 < \max V_{cp} \leq 40$: middelmatig verschil.
- $\max V_{cp} \leq 20$: gering verschil.

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

- $E > 0{,}8$: groot verschil.
- $0{,}4 < E \leq 0{,}8$: middelmatig verschil.
- $E \leq 0{,}4$: gering verschil.

Twee boxplots vergelijken

- De boxen[1] overlappen elkaar niet: groot verschil.
- De boxen overlappen elkaar wel en de mediaan van de ene boxplot ligt buiten de box van de andere boxplot: middelmatig verschil.
- Alle andere gevallen: gering verschil.

[1] De 'box' is het interval vanaf het eerste kwartiel tot en met het derde kwartiel.

§ 4.2 Doel van deze module

De leerdoelen van deze module zijn:

- Hoe kun je op basis van een steekproef uitspraken doen over:
    - Een proportie in een populatie en de betrouwbaarheid ervan?
    - Een gemiddelde in een populatie en de betrouwbaarheid ervan?
- Hoe kun je op basis van steekproefgegevens of populatiegegevens een uitspraak doen over de omvang van het verschil tussen twee groepen?
- Hoe kun je op basis van steekproefgegevens of populatiegegevens een uitspraak doen over de samenhang tussen twee kwantitatieve variabelen?
- Hoe pas je statistiek toe op grote gegevensbestanden met behulp van ICT?

De opbouw van de module is als volgt:

Paragrafen 4.3 en 4.4 behandelen hoe je op basis van een steekproef uitspraken kunt doen over een populatieproportie en -gemiddelde.
Paragraaf 4.5 behandelt hoe je uitspraken kunt doen over de omvang van het verschil tussen twee groepen.
Paragraaf 4.6 behandelt hoe je uitspraken kunt doen over de samenhang tussen twee kwantitatieve variabelen.
Paragraaf 4.7 bevat gemengde opdrachten.
Paragraaf 4.8 geeft een terugblik op paragraaf 4.3 tot en met 4.7.
De paragrafen 4.3 tot en met 4.8 dekken de betreffende eindtermen van het centraal examen.
Paragraaf 4.9 bevat een lessenserie die gericht is op het als vierde genoemde doel van deze module: het toepassen van statistiek op grote gegevensbestanden met behulp van ICT. Deze lessenserie kan – eventueel in een aangepaste vorm – gegeven worden als een praktische opdracht in het kader van het schoolexamen.
In paragraaf 4.10 staat een diagnostische computertoets. Een soortgelijke toets kan afgenomen worden in het kader van het schoolexamen.
De paragrafen 4.9 en 4.10 geven dus een mogelijke invulling van de statistiekonderdelen die vallen onder het schoolexamen.

§ 4.3 Populatieproportie

§ 4.3.1 Introductie

- Ga naar www.benikgemiddeld.nl.
- Klik op een persoon, voer je leeftijd in en bevestig dat jij dit bent.
- Kies voor ‘Gezondheid’ en daarna voor ‘Drinken’.
- Kies voor 12 tot en met 18 jaar.
- Lees nu het percentage drinkers af.
- Voor jongens is dit 35 procent.

Hier staat dus dat 35 procent van de jongens van 12 tot en met 18 jaar in Nederland wel eens alcohol drinkt. Dit is een uitspraak over een populatieproportie.
Andere voorbeelden van uitspraken over een populatieproportie zijn:

- 75 procent van de Nederlanders checkt elke 3 tot 6 minuten zijn smartphone.
- Ruim 70 procent van de Nederlanders geeft minder geld uit aan kleding, vakantie en vrije tijd.

In deze paragraaf gaan we nader in op uitspraken over een populatieproportie. Beide uitspraken hierboven gaan over de populatie Nederlanders. Het mag duidelijk zijn dat beide uitspraken gebaseerd zijn op een steekproef uit de populatie. We veronderstellen steeds dat de steekproef aselect is.
In feite wordt hier nu de steekproefproportie gepresenteerd als de meest aannemelijke schatting voor de populatieproportie. De betrouwbaarheid van deze schatting hangt af van de omvang van de steekproef. In deze paragraaf gaan we dan ook met name in op de invloed van de omvang van de steekproef op de betrouwbaarheid van de schatting van de populatieproportie.

§ 4.3.2 Centrale vraag

Stel dat de bovenstaande uitspraak over het checken van de smartphone gebaseerd is op een steekproef van 200 mensen.
In de steekproef geven 150 mensen aan dat ze elke 3 tot 6 minuten hun smartphone checken.

Hoe doe je op basis van deze gegevens een uitspraak over de proportie Nederlanders dat elke 3 tot 6 minuten zijn smartphone checkt en de betrouwbaarheid hiervan?

Het meest waarschijnlijk is dat de proportie Nederlanders dat elke 3 tot 6 minuten zijn smartphone checkt hetzelfde is als de proportie in steekproef. Dus de meest aannemelijke schatting van de populatieproportie is gelijk aan de steekproefproportie. In dit geval dus 150 / 200 = 0,75.
De populatieproportie hoeft natuurlijk niet precies gelijk te zijn aan 0,75. Deze zou ook wel iets groter of iets kleiner kunnen zijn. Het is echter meer waarschijnlijk dat de populatieproportie gelijk is aan 0,70 dan aan 0,55, omdat 0,7 nu eenmaal dichter in de buurt ligt van de steekproefproportie van 0,75.

§ 4.3.3 Antwoord op de centrale vraag

Er bestaat een eenvoudige vuistregel waarmee we een uitspraak kunnen doen over de populatieproportie en de betrouwbaarheid ervan.
Deze vuistregel luidt: met 95 procent zekerheid ligt de populatieproportie tussen de volgende grenzen:

$$p \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$$

met p de steekproefproportie en n de steekproefomvang.

Dus als er in een steekproef van 200 waarnemingen een steekproefproportie van 0,75 gevonden wordt, dan zal met 95 procent zekerheid de populatieproportie liggen tussen de grenzen $0{,}75 - 2 \cdot \sqrt{\frac{0{,}75 \cdot 0{,}25}{200}}$ en $0{,}75 + 2 \cdot \sqrt{\frac{0{,}75 \cdot 0{,}25}{200}}$.

De linkergrens (ondergrens) is dus 0,69 en de rechtergrens (bovengrens) is dus 0,81. Samen vormen deze grenzen dus het interval [0,69; 0,81]. We noemen dit interval het 95%-betrouwbaarheidsinterval voor de populatieproportie.

EXTRA (niet in de eindtermen)
De formule is gebaseerd op de centrale limietstelling die zegt dat de verdeling van de steekproefproportie voor voldoende grote steekproeven benaderd kan worden door een normale verdeling.
Bij een normale verdeling geldt de vuistregel dat 95 procent van alle waarnemingen zich zal bevinden tussen het gemiddelde plus of min twee keer de standaardafwijking. De 2 in de formule is een afgeronde waarde van de zogenaamde z-waarde die hoort bij 95 procent. Een preciezere waarde is 1,96.
Wanneer er voor een andere betrouwbaarheid wordt gekozen dan 95 procent, dan moet de 2 in de formule vervangen worden door de z-waarde die hoort bij die andere betrouwbaarheid. Als er voor meer betrouwbaarheid wordt gekozen, bijvoorbeeld 99 procent, dan hoort daar de z-waarde 2,58 bij. Als je dus in de formule de 2 vervangt door 2,58, dan krijg je het 99%-betrouwbaarheidsinterval.

Ander voorbeeld: kies je voor minder betrouwbaarheid, bijvoorbeeld 90 procent, dan hoort daar de z-waarde 1,65 bij. Als je dus in de formule de 2 vervangt door 1,65, dan krijg je het 90%-betrouwbaarheidsinterval.

§ 4.3.4 Oefenen

Opgave 1
Aan het begin van deze paragraaf staat de uitspraak:
Ruim 70 procent van de Nederlanders geeft minder geld uit aan kleding, vakantie en vrije tijd.

Veronderstel dat deze uitspraak is gebaseerd op een aselecte steekproef van 400 Nederlanders.
Bereken op basis van deze gegevens een 95%-betrouwbaarheidsinterval voor de proportie Nederlanders dat minder geld uitgeeft aan kleding, vakantie en vrije tijd.

Opgave 2
Hieronder zie je een frequentietabel gemaakt met de gegevens van een aselecte steekproef onder leerlingen uit leerjaar 1 en 2 van het voortgezet onderwijs.
Bereken het 95%-betrouwbaarheidsinterval voor de proportie leerlingen waarvan wiskunde het favoriete vak is.

Vak	**Freq.**	**Perc.**
Nederlands	1233	2,46
Engels	2352	4,70
Frans	1831	3,66
Duits	835	1,67
Geschiedenis	1879	3,75
Aardrijkskunde	772	1,54
Wiskunde	4653	9,29
Natuur- en scheikunde	917	1,83
Biologie	2008	4,01
Economie	463	0,92
Techniek	3939	7,87
Verzorging	1580	3,16
Informatiekunde	3104	6,20
Lichamelijke opvoeding	13988	27,94
Beeldende vorming	2406	4,81
Muziek	3713	7,42
Dans	427	0,85
Drama	907	1,81
Ander vak	3064	6,12
Totaal	50071	100%

Opgave 3
Bij een onderzoek naar de slagingskans voor het rijexamen wordt van een aselecte steekproef van 800 pogingen vastgesteld of het examen is gehaald of niet. Van die 800 pogingen blijken er 683 succesvol te zijn.

a. Worden de 90%-, 95%- en 99%-betrouwbaarheidsintervallen voor de slagingskans voor het rijexamen steeds breder of smaller? Licht je antwoord toe.
b. Met behulp van EXTRA (zie § 4.3.3) kun je die intervallen ook berekenen. Doe dat.

§ 4.3.5 Om te onthouden

Als je op basis van een steekproef een uitspraak wilt doen over een populatieproportie en de betrouwbaarheid ervan, dan kun je gebruikmaken van een eenvoudige vuistregel.

Deze vuistregel zegt dat bij een aselecte steekproef geldt dat met 95 procent zekerheid de populatieproportie tussen de volgende grenzen ligt:

$$p \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$$

met p de steekproefproportie en n de steekproefomvang.

Dit noemen we het 95%-betrouwbaarheidsinterval voor de populatieproportie.

§ 4.3.6 Geïntegreerd oefenen

Opgave 4

Het begrip ‘Auschwitz’ zegt een op de vijf Duitse jongeren niets. Uit een opiniepeiling van het Duitse tijdschrift **STERN** blijkt dat slechts 79 procent van de jongeren weet dat Auschwitz een voormalig nazivernietigingskamp is. **STERN** hield de opiniepeiling in aanloop naar de Internationale Herdenkingsdag voor de Holocaust.

a. Wat is de populatie waarop de opiniepeiling betrekking heeft?

Veronderstel dat de opiniepeiling is gehouden onder een aselecte steekproef van 360 personen uit de populatie.

b. Bereken op basis van deze gegevens een 95%-betrouwbaarheidsinterval voor de populatieproportie.
c. Wat zouden de houders van de opiniepeiling kunnen doen om ons ervan te overtuigen dat de steekproef aselect is?
d. Noem enkele situaties waarin er aan getwijfeld mag worden dat de steekproef aselect is.

Opgave 5

Ouderen nemen vaak te veel of juist te weinig medicijnen. Dat kan leiden tot vervelende bijwerkingen en zelfs onnodige ziekenhuisopnamen, zo blijkt uit een onderzoek van het RIVM (Rijksinstituut voor Volksgezondheid en Milieu).

Bijna 20 procent van de 75-plussers neemt dagelijks negen of meer geneesmiddelen. Het probleem met de dosering blijkt vaak te ontstaan door gebrekkige informatietechnologie, waardoor communicatie niet altijd vlot verloopt. Ook weten artsen van elkaar vaak niet welke medicijnen zij voorschrijven.

a. Wat is de populatie waarop dit bericht betrekking heeft?

Veronderstel dat het bericht is gebaseerd op een aselecte steekproef van 1000 personen uit de populatie.

b. Bereken het 95%-betrouwbaarheidsinterval voor de populatieproportie.

We laten nu de veronderstelling dat de steekproef uit 1000 personen bestaat los en we veronderstellen dat de onderzoekers met 95 procent zekerheid de populatieproportie willen schatten op 2 decimalen nauwkeurig.

c. Bereken hoe groot de steekproef zal moeten zijn om aan deze eis te voldoen.

Opgave 6

Veel fruittelers houden zich niet aan de regels. Dat stelt de Nederlandse Voedsel- en Warenautoriteit (NVWA) na de jaarlijkse controle onder 100 van de 3200 telers. Ongeveer een derde van hen gebruikt verboden middelen om fruitgewassen te beschermen tegen insecten en ziektes. Milieudefensie noemt het ‘ongehoord’ dat een derde van de telers verboden gif gebruikt.

a. Wat is hier de populatie?

Veronderstel dat de 100 onderzochte telers een aselecte steekproef vormen uit de populatie.

b. Bereken het 95%-betrouwbaarheidsinterval.
c. Op welke wijze zou de NVWA kunnen garanderen dat de steekproef aselect is?
d. Wat zouden redenen kunnen zijn dat de steekproef niet aselect is?

§ 4.4 Populatiegemiddelde

§ 4.4.1 Introductie

- Ga naar www.benikgemiddeld.nl.
- Klik op een persoon, voer je leeftijd in en bevestig dat jij dit bent.
- Kies voor ‘Gezondheid’ en daarna voor ‘Drinken’.
- Kies voor 12 tot en met 18 jaar.
- Lees nu het gemiddelde aantal alcoholconsumpties af.
- Voor jongens is dit 7,1.

Dit is een uitspraak over een populatiegemiddelde.
Andere voorbeelden van uitspraken over een populatiegemiddelde zijn:

- We staren gemiddeld 42 minuten per dag naar het beeldscherm van onze telefoon.
- De ruim 70 procent die bezuinigt op kleding, vrije tijd en boodschappen, spaart daarmee al meer dan 200 euro per maand uit. Aan kleding wordt elke maand 53 euro minder uitgegeven, aan vrije tijd 103 euro en aan dagelijkse boodschappen 60 euro.

In deze paragraaf gaan we nader in op uitspraken over een populatiegemiddelde op basis van een steekproef en de betrouwbaarheid ervan. We veronderstellen steeds dat de steekproef aselect is, dat van de steekproef de omvang bekend is en dat in de steekproef het gemiddelde en de standaardafwijking zijn berekend. Met deze steekproefresultaten doen we een uitspraak over het populatiegemiddelde en de betrouwbaarheid ervan.

§ 4.4.2 Centrale vraag

Stel dat bovenstaande uitspraak over het staren naar het beeldscherm van onze smartphone gebaseerd is op een steekproef van 100 mensen, dat het steekproefgemiddelde gelijk is aan 42 minuten en dat de standaardafwijking in de steekproef gelijk is aan 8 minuten.
Hoe kun je op basis van deze steekproefresultaten een uitspraak doen over het populatiegemiddelde en de betrouwbaarheid ervan?

Het meest aannemelijk is dat het populatiegemiddelde gelijk is aan 42 minuten, maar dat hoeft natuurlijk niet precies te kloppen. Het populatiegemiddelde zal waarschijnlijk in de buurt liggen van die 42 minuten. Het is waarschijnlijker dat het populatiegemiddelde gelijk is aan 40 minuten dan 36 minuten, omdat 40 dichter bij het steekproefgemiddelde ligt.

§ 4.4.3 Antwoord op de centrale vraag

Als van een steekproef de omvang, het gemiddelde en de standaardafwijking bekend zijn, dan kun je het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde berekenen als:

$$steekproef gemiddelde \ \pm 2 \cdot \frac{steekproef standaardafwijking}{\sqrt{steekproef omvang}}$$

In bovenstaande situatie is het steekproefgemiddelde 42 minuten, de standaardafwijking 8 minuten en de steekproefomvang 100. Invullen van deze gegevens levert een ondergrens van $42 - 2 \cdot \frac{8}{\sqrt{100}}$ en een bovengrens van $42 + 2 \cdot \frac{8}{\sqrt{100}}$. Dus de ondergrens is 40,4 en de bovengrens is 43,6.

Met 95 procent zekerheid ligt het populatiegemiddelde dus tussen 40,4 en 43,6. Met 95 procent zekerheid kunnen we dus zeggen dat mensen tussen de 40,4 en 43,6 minuten per dag naar het beeldscherm van hun smartphone staren.

In formule

Het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde is $\overline{X} \pm 2 \cdot \frac{S}{\sqrt{n}}$

met $\overline{X}$ het steekproefgemiddelde, n de steekproefomvang en S de steekproefstandaardafwijking.

EXTRA (niet in eindtermen)

De formule is gebaseerd op de centrale limietstelling die zegt dat de verdeling van het steekproefgemiddelde voor voldoende grote steekproeven benaderd kan worden door een normale verdeling.

Bij een normale verdeling geldt de vuistregel dat 95 procent van alle waarnemingen zich zal bevinden tussen het gemiddelde plus of min twee keer de standaardafwijking. Dit is precies de opbouw van de formule.

Het gemiddelde is $\overline{X}$ en de standaardafwijking is $\frac{S}{\sqrt{n}}$.

De 2 in de formule is een afgeronde waarde van de zogenaamde z-waarde die hoort bij 95 procent (1,96).

Net als bij het betrouwbaarheidsinterval voor de populatieproportie, kun je hier weer voor een andere betrouwbaarheid kiezen als je de waarde van 2 vervangt door de z-waarde die hoort bij de gekozen betrouwbaarheid.

§ 4.4.4 Oefenen

Opgave 7

Uit een aselecte steekproef van ruim 50.000 leerlingen uit leerjaar 1 en 2 van het voortgezet onderwijs zijn onderstaande kentallen bekend van het aantal uren dat per week ze naar sport kijken.

Variabele	**SPORT**
Aantal waarnemingen	50071
Gemiddelde	4,4
Mediaan	4
Modus	2
Minimum	0
Maximum	70
SDn-1	4,65
SDn	4,65
VARn-1	21,58
VARn	21,58

a. Bereken het 95%-betrouwbaarheidsinterval van het aantal uren dat ze naar sport kijken.

Ook zijn onderstaande kentallen berekend van het zakgeld dat ze per week krijgen.

Variabele	**ZAKGELD**
Aantal waarnemingen	50071
Gemiddelde	11,0
Mediaan	8
Modus	10
Minimum	0
Maximum	500
SDn-1	19,80
SDn	19,80
VARn-1	391,91
VARn	391,91

b. Bereken het 95%-betrouwbaarheidsinterval van het zakgeld dat ze per week krijgen.

Opgave 8

In een aselecte steekproef onder Nederlanders blijkt dat 70 procent bezuinigt op kleding, vrije tijd en boodschappen. Deze groep bespaart daarmee gemiddeld genomen 200 euro per maand.
Veronderstel dat de overige 30 procent niet bezuinigt op kleding, vrije tijd en boodschappen.

a. Bereken de gemiddelde besparing per maand op kleding, vrije tijd en boodschappen in de steekproef.

Stel dat de steekproefomvang 800 is en de standaardafwijking in de steekproef gelijk is aan 120.

b. Bereken het 95%-betrouwbaarheidsinterval voor de gemiddelde besparing op kleding, vrije tijd en boodschappen.

§ 4.4.5 Om te onthouden

Als van een steekproef de omvang, het gemiddelde en de standaardafwijking bekend zijn, dan kun je het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde berekenen als:

$$\overline{X} \pm 2 \cdot \frac{S}{\sqrt{n}}$$

met $\overline{X}$ het steekproefgemiddelde, n de steekproefomvang en S de steekproefstandaardafwijking.

§ 4.4.6 Geïntegreerd oefenen

Opgave 9

Kinderen in een mammoetklas met meer dan 30 leerlingen, omdat de school moet bezuinigen? Niet alleen basisscholen zijn er ongelukkig mee. Ouders vrezen dat hun kroost te weinig aandacht krijgt. "*We draaien de klok terug naar begin jaren negentig*", stelt pedagoog Bas Levering. Tussen 1997 en 2002 daalde het gemiddeld aantal kinderen in een onderbouwklas van 23,7 naar 20,9 en kwamen er onderwijsassistenten. Uit een evaluatierapport blijkt dat kleinere groepen en meer handen in de klas de kwaliteit van het onderwijs in de onderbouw van het basisonderwijs hebben verbeterd.

a. Wat is hier de populatie?

Veronderstel dat de gegevens over 1997 gebaseerd zijn op een aselecte steekproef van 80 klassen en dat de standaardafwijking gelijk is aan 3.

b. Bereken het 95%-betrouwbaarheidsinterval voor de gemiddelde omvang van een onderbouwklas in 1997.

Veronderstel dat de gegevens over 2002 gebaseerd zijn op een aselecte steekproef van 120 klassen en dat de standaardafwijking gelijk is aan 4.

c. Bereken het 95%-betrouwbaarheidsinterval voor de gemiddelde omvang van een onderbouwklas in 2002. Overlapt dit 95% betrouwbaarheidsinterval met dat van 1997?

Veronderstel dat een onderzoeker met 95 procent zekerheid de gemiddelde omvang van een onderbouwklas op 1 decimaal nauwkeurig wil vaststellen.
Veronderstel verder dat de standaardafwijking gelijk is aan 4.

d. Bereken hoe groot de steekproef minimaal moet zijn om aan de gegeven eis te kunnen voldoen.

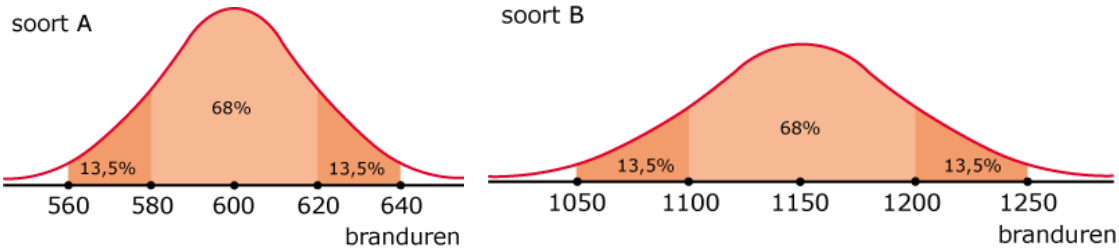
Het is niet noodzakelijk dat een onderzoek naar de omvang van klassen in het basisonderwijs zich baseert op steekproefgegevens.

e. Welke organisatie zou voor zo'n onderzoek kunnen beschikken over populatiegegevens?

Opgave 10

Van lampen van soort A is de levensduur van 500 aselect gekozen exemplaren gemeten.
Van soort B zijn er aselect 1200 lampen gekozen.
Het aantal branduren blijkt in beide gevallen vrijwel normaal verdeeld te zijn.

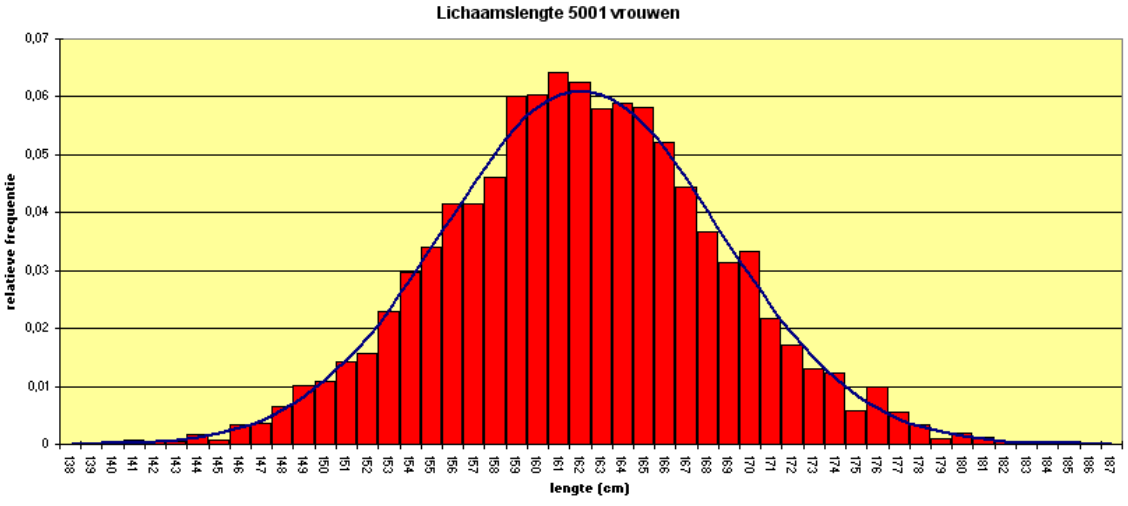
Hieronder zie je een schets van de steekproefverdelingen. Enkele percentages zijn gegeven om onder andere de standaardafwijkingen te kunnen bepalen.



Stel op basis van deze gegevens de 95%-betrouwbaarheidsintervallen op voor de levensduur van soort A en soort B.

Opgave 11

In 1947 onderzochten Freudenthal en Sittig 5001 vrouwen in opdracht van het warenhuis De Bijenkorf met als doel het ontwerpen van een maatsysteem voor kleding. Van deze vrouwen werd onder andere de lichaamslengte in centimeters gemeten.
Hieronder zie je de verdeling van de lichaamslengtes van 5001 aselect gekozen vrouwen uit het onderzoek van Freudenthal en Sittig nog eens.



Maak met behulp van deze gegevens een 95%-betrouwbaarheidsinterval voor de lengte van de vrouwen destijds.

Opgave 12

Uit de registratie van medische gegevens is bekend dat de duur in dagen van de menselijke zwangerschap vrijwel normaal verdeeld is met gemiddeld 266 dagen en standaardafwijking 16.
Een gynaecoloog vraagt zich af wat de gemiddelde duur van de zwangerschap is van vrouwen die bij hem op medische indicatie in het ziekenhuis bevallen. Om hiervoor een betrouwbaarheidsinterval te berekenen, baseert hij zich op een aselecte steekproef van 64 vrouwen die bij hem op medische indicatie in het ziekenhuis zijn bevallen. De gemiddelde zwangerschapsduur van deze vrouwen is 250 dagen en de standaardafwijking is 15.
Laat met geschikte berekeningen zien dat de waarde van 266 dagen niet ligt in het 95%-betrouwbaarheidsinterval.

Begrepen?

Om te controleren of je de inhoud van paragraaf 3 en 4 hebt begrepen doe je onderstaande oefening:

- Ga naar www.statslc.com/videos.
- Kies bij '*Confidence Interval for a Mean'* voor het filmpje '*Understanding Confidence Intervals: Statistics Help'*.
- Geef een samenvatting van dit filmpje.

§ 4.5 Verschil tussen twee groepen

Ga naar www.youtube.com en bekijk het filmpje '*PVV miep snapt statistiek niet*'.
In het filmpje gaat het over twee groepen die in aanraking zijn gekomen met justitie voor een licht vergrijp. De ene groep heeft een celstraf gekregen en de andere groep heeft een taakstraf gehad. Voor beide groepen is gekeken naar het percentage dat recidiveert. Dat wil zeggen: het percentage dat voor hetzelfde vergrijp opnieuw in aanraking komt met justitie. In de groep die een taakstraf heeft gehad is het percentage recidivisten (aanzienlijk) lager dan in de groep die een celstraf heeft gehad. Wat kun je nu zeggen over het verschil tussen de twee groepen? Dat is het onderwerp van deze paragraaf.

Andere voorbeelden van verschillen tussen twee groepen zijn:

- Ruim de helft van de mannen en bijna een derde van de vrouwen zegt wel eens zo dronken te zijn geweest dat zij niet meer wisten hoe zij thuis zijn gekomen. Dit blijkt uit onderzoek van de drankenproducent Diageo. Diageo heeft een filmpje gemaakt waarin feestgangers dronken over straat zwalken. De video hoort bij de campagne *Think how you drink* waarbij mensen hun eigen drankgebruik in kaart kunnen brengen.
- Europa raakt hopeloos achterop met zijn 4G-netwerken voor mobiel- en dataverkeer. Terwijl in Amerika al 90 procent zo'n supersnelle aansluiting heeft, zit in Europa drie kwart nog zonder.

Hoe je het verschil tussen twee groepen kwantificeert hangt af van het meetniveau van de variabele die je wilt bekijken. Eerst vergelijken we twee groepen op een nominale variabele. Daarna vergelijken we twee groepen op een ordinale variabele en tot slot op een kwantitatieve variabele.

§ 4.5.1 Op een nominale variabele (phi)

§ 4.5.1.1 Introductie

Vaak worden gegevens van twee groepen op één nominale variabele met twee mogelijke uitkomsten gepresenteerd in een 2x2-tabel. Hieronder staat een voorbeeld. In deze 2x2-kruistabel zijn de variabelen geslacht (J = jongen, M = meisje) en wiskundegroep met elkaar gecombineerd.

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	13	30	43
B	56	55	111
Totaal	**69**	85	154

§ 4.5.1.2 Centrale vraag

Is er in de kruistabel sprake van een groot, middelmatig of gering verschil tussen jongens en meisjes met betrekking de keuze van wiskunde B?

Opgave 13

a. Bedenk zelf enkele 2x2-tabellen waarbij duidelijk sprake is van een gering verschil tussen jongens en meisjes voor wat betreft de keuze van wiskunde B.
Met andere woorden: jongens en meisjes kiezen op ongeveer dezelfde wijze.
Vul daarvoor volgende lege 2x2-tabellen in. Zorg dat je iedere keer uitkomt op 100 leerlingen.
Licht je antwoord toe.

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A			
B			
Totaal			100

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A			
B			
Totaal			100

b. Bedenk nu ook enkele 2x2-tabellen waarbij er een grote samenhang is tussen de variabelen geslacht en wiskundegroep. Het verschil tussen de wijze waarop jongens en meisjes kiezen is nu groot.
Licht je antwoord toe.

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A			
B			
Totaal			100

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A			
B			
Totaal			100

§ 4.5.1.3 Antwoord op de centrale vraag

Een maat die veel gebruikt wordt om de verschillen tussen variabelen te meten in een 2x2 tabel is phi. Op het formuleblad staan vuistregels voor phi.

Stel: 2×2-kruistabel $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Dan berekenen we:

$$phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

- $phi < -0{,}4$ of $phi > 0{,}4$: groot verschil.
- $-0{,}4 < phi < -0{,}2$ of $0{,}2 < phi < 0{,}4$: middelmatig verschil.
- $-0{,}2 < phi < 0{,}2$: gering verschil.

Opgave 14

Test je voorbeelden uit opgave 2 met voorgaande formule en vuistregels.

Opgave 15

Bekijk de volgende 2x2-tabellen goed en voorspel of er sprake is van een groot, middelmatig of gering verschil. Bereken daarna phi en controleer je voorspelling met behulp van de vuistregels.

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	10	30	
B	15	45	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	3	1	
B	12	4	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	3	0	
B	0	1	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	12	0	
B	0	4	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	12	1	
B	1	4	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	12	1	
B	0	5	
Totaal			

Opgave 16
Bij de volgende 2x2-tabellen is het lastiger te voorspellen. Probeer toch een voorspelling te doen en bereken daarna weer phi en controleer je voorspelling met behulp van de vuistregels.

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	7	1	
B	12	4	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	3	1	
B	12	3	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	12	5	
B	30	14	
Totaal			

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	36	15	
B	80	34	
Totaal			

Opgave 17
Bekijk opnieuw de centrale vraag van deze paragraaf.

Bepaal met behulp van phi of er in de tabel hiernaast sprake is van een groot, middelmatig of gering verschil.

	Geslacht		
Wiskunde-groep	**J**	**M**	**Totaal**
A	13	30	43
B	56	55	111
Totaal	69	85	154

§ 4.5.1.4 Oefenen

Opgave 18
Er is onderzoek gedaan naar het favoriete avondje uit onder jongeren.
Zie de tabel hiernaast:

	Jongens	**Meisjes**
Film	745	667
Disco	580	370

Is er verschil tussen jongens en meisjes voor wat betreft hun voorkeur voor een avondje uit?

Opgave 19
Bij een onderzoek over kleurenblindheid is 1000 mensen gevraagd of ze een vorm van kleurenblindheid hebben. In totaal worden 600 mannen bevraagd, waarvan er 65 aangeven kleurenblind te zijn. Van de vrouwen blijken er maar 7 kleurenblind te zijn.
Maak een 2x2-tabel en toon aan dat er een gering verschil is tussen mannen en vrouwen.

§ 4.5.1.5. Om te onthouden

Om te bepalen of er sprake is van een (groot) verschil tussen twee groepen op een nominale variabele bereken je phi. Met behulp van vuistregels geef je een oordeel over de omvang van het verschil tussen de twee groepen.

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

- $phi < -0{,}4$ of $phi > 0{,}4$: groot verschil.
- $-0{,}4 < phi < -0{,}2$ of $0{,}2 < phi < 0{,}4$: middelmatig verschil.
- $-0{,}2 < phi < 0{,}2$: gering verschil.

§ 4.5.1.6 Geïntegreerd oefenen

Opgave 20

Gegeven is onderstaande kruistabel voor de variabelen *ontbeten* en *geslacht*.

	GESLACHT		
ONTBETEN	**Jongen**	**Meisje**	**Totaal**
Ja	21888	21826	43714
Iets meegenomen	1387	2105	3492
Nee	1197	1668	2865
Totaal	24472	25599	50071

Onderzoek hoe groot het verschil is tussen jongens en meisjes voor wat betreft het percentage dat niet heeft ontbeten.

Opgave 21

Gegeven is onderstaande kruistabel van favoriete vak en leerjaar.

	Leerjaar		
Vak	**1**	**2**	**Totaal**
Nederlands	815	418	1233
Engels	1311	1041	2352
Frans	1284	547	1831
Duits	183	652	835
Geschiedenis	886	993	1879
Aardrijkskunde	351	421	772
Wiskunde	3443	1210	4653
Natuur- en scheikunde	173	744	917
Biologie	1202	806	2008
Economie	20	443	463
Techniek	2403	1536	3939
Verzorging	568	1012	1580
Informatiekunde	2273	831	3104
Lichamelijke opvoeding	6342	7646	13988
Beeldende vorming	1176	1230	2406
Muziek	2270	1443	3713
Dans	214	213	427
Drama	474	433	907
Ander vak	1693	1371	3064
Totaal	27081	22990	50071

Onderzoek hoe groot het verschil is tussen leerjaar 1 en 2 voor wat betreft het aantal leerlingen dat wiskunde als favoriete vak heeft.

§ 4.5.2 Op een ordinale variabele ($\max V_{cp}$)

§ 4.5.2.1 Introductie

Havo-4 leerlingen bezoeken een toneelvoorstelling. Daarna vraagt men hoe ze deze voorstelling hebben beleefd. In onderstaande tabel zie je de resultaten uitgesplitst naar profiel.

	CM	EM	NG	NT
1 = niet boeiend	5	8	6	17
2 = gaat wel	12	12	18	13
3 = boeiend	9	18	15	8
4 = erg boeiend	9	10	6	2

§ 4.5.2.2 Centrale vraag

Hoe kun je aan de hand van de tabel bepalen of er sprake is van een (groot) verschil in de voorstellingsbeleving van EM-leerlingen ten opzichte van NG-leerlingen?

§ 4.5.2.3 Antwoord op de centrale vraag

Om deze vraag te beantwoorden, bereken je het zogenaamde maximale cumulatieve percentageverschil uit. Hieronder staat hoe je dit doet.

Eerst moet je de aantallen omzetten naar percentages. In de volgende tabel zie je onder ‘p’ de percentages en onder ‘cp’ de cumulatieve percentages.

	EM			NG			
	aantal	p	cp	aantal	p	cp	Vcp
1 = niet boeiend	8	16,7	16,7	6	13,3	13,3	3,3
2 = gaat wel	12	25,0	41,7	18	40,0	53,3	11,7
3 = boeiend	18	37,5	79,2	15	33,3	86,7	7,5
4 = erg boeiend	10	20,8	100,0	6	13,3	100,0	0,0
	48			45			

Het valt op dat de percentages van de EM-leerlingen en de NG-leerlingen nogal verschillen.
Veel meer NG-leerlingen dan EM-leerlingen hebben “*gaat wel*” geantwoord.
En veel minder NG-leerlingen dan EM-leerlingen hebben “*erg boeiend*” geantwoord.
Toch hebben ook minder NG-leerlingen “*niet boeiend*” geantwoord.

Vervolgens kijk je naar de verschillen in cumulatieve percentages, waarbij je bij negatieve verschillen het minteken weglaat. Deze verschillen in cumulatieve percentages staan in de laatste kolom (Vcp).

Tot slot kijk je naar de grootste waarde in deze kolom. Dit noemen we het maximale cumulatieve percentageverschil. In ons voorbeeld is dit dus 11,7.

Merk op dat de steekproefomvang in dit voorbeeld 93 is. Op het formuleblad staat dat het maximale cumulatieve percentageverschil toegepast mag worden als de steekproefomvang groter is dan 100.
Dit voorbeeld voldoet dus (net) niet aan deze eis. Helaas was deze eis ten tijde van het schrijven van dit materiaal nog niet bekend bij de auteurs.

Nu je de waarde van het maximale cumulatieve percentageverschil hebt berekend, verbind je er een oordeel aan. Betekent de gevonden waarde van 11,7 nu dat het verschil klein is of juist groot?

Voor het geven van zo'n oordeel maak je weer gebruik van vuistregels. Als vuistregels spreken we af:

Maximale verschil in cumulatief percentage ($\max V_{cp}$)

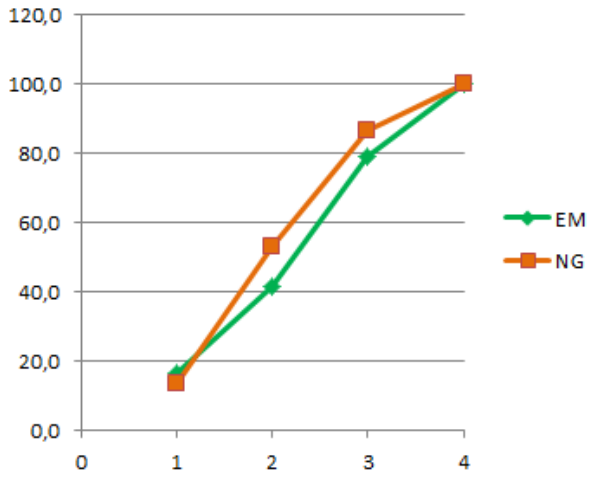
- $\max V_{cp} > 40$: groot verschil.
- $20 < \max V_{cp} \leq 40$: middelmatig verschil.
- $\max V_{cp} \leq 20$: gering verschil.

In ons voorbeeld is het maximale cumulatieve percentageverschil gelijk aan 11,7,
dus op basis van de vuistregels is het verschil tussen de twee groepen gering.
Met andere woorden: het verschil in de voorstellingsbeleving tussen EM- en NG-leerlingen is gering.

We hebben de centrale vraag beantwoord door gebruik te maken van een tabel.
Je kunt het antwoord ook vinden met behulp van een grafiek.

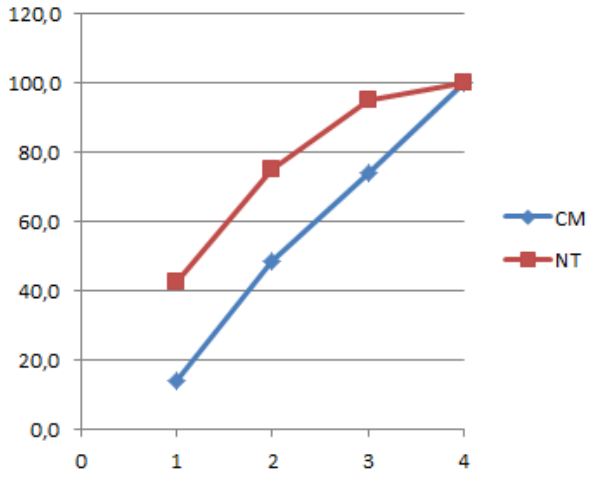
Dat doe je als volgt

De cumulatieve percentages kun je uitzetten in cumulatieve relatieve frequentiepolygonen. Hier zie je dergelijke polygonen bij het vergelijken van de EM- en de NG-groep en van de CM- en NT-groep.



Op de x-as staat de voorstellingsbeleving en op de y-as het cumulatieve percentage van de leerlingen. Om uit zo'n figuur het maximale cumulatieve percentageverschil af te lezen, moet je kijken waar het verschil (in verticale zin) tussen beide polygonen het grootst is. In de figuur kun je zien dat dit is bij voorstellingsbeleving 2 (op de x-as). Het verschil in cumulatief percentage is daar gelijk aan ongeveer 12 procent (aflezen op de y-as).
Dit komt goed overeen met de hierboven berekende 11,7 procent met behulp van de tabel. De conclusie luidt dan ook hetzelfde: het verschil tussen EM- en NG-leerlingen in voorstellingsbeleving is gering.

Tot slot kijken we naar het verschil in voorstellingsbeleving tussen CM- en NT-leerlingen.
We maken hierbij gebruik van de grafiek.



In deze figuur liggen de polygonen duidelijk verder uit elkaar dan in de bovenstaande figuur voor EM- versus NG-leerlingen.
Dit betekent dat het verschil tussen CM- en NT-leerlingen voor wat betreft hun voorstellingsbeleving groter is dan het verschil tussen de EM- en de NG-leerlingen. In de figuur voor CM-leerlingen versus NT-leerlingen zit het maximale cumulatieve percentageverschil bij voorstellingsbeleving 1 (op de x-as). Het verschil in cumulatief percentage is daar ongeveer 26 procent.
Op basis van de bovenstaande vuistregel kunnen we dan zeggen dat het verschil in voorstellingsbeleving tussen CM- en NT-leerlingen middelmatig is.

§ 4.5.2.4 Oefenen

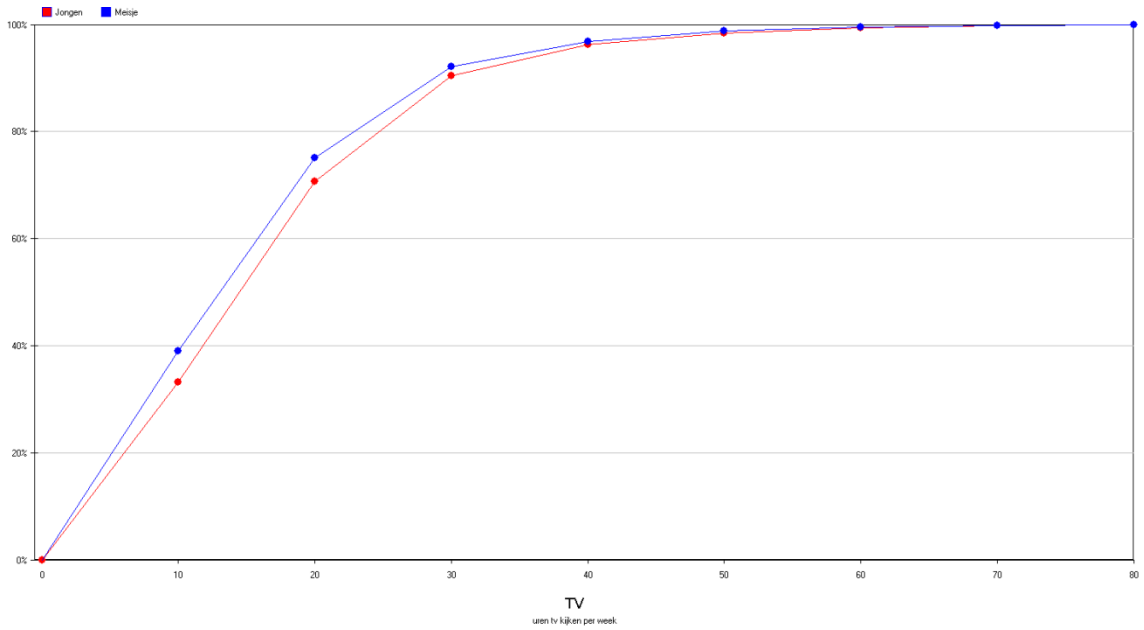
Opgave 22
Gegeven is onderstaande frequentietabel voor de lengte van jongens respectievelijk meisjes.

Geslacht	**Jongen**				**Meisje**				
Lengte	**Freq.**	**Perc.**	**Cumul.**	**Cumul. %**	**Freq.**	**Perc.**	**Cumul.**	**Cumul. %**	**Totaal**
130-139	1	0,00	1	0,00	0	0,00	0	0,00	1
140-149	1106	4,52	1107	4,52	927	3,62	927	3,62	2033
150-159	6396	26,14	7503	30,66	6630	25,90	7557	29,52	13.026
160-169	10.047	41,06	17.550	71,71	12.989	50,74	20.546	80,26	23.036
170-179	5481	22,40	23.031	94,11	4791	18,72	25.337	98,98	10.272
180-189	1278	5,22	24.309	99,33	218	0,85	25.555	99,83	1496
190-199	115	0,47	24.424	99,80	24	0,09	25.579	99,92	139
200-209	33	0,13	24.457	99,94	15	0,06	25.594	99,98	48
210-219	15	0,06	24.472	100,00	5	0,02	25.599	100,00	20
Totaal	24.472	100%	24.472	100%	25.599	100%	25599	100%	50.071

Onderzoek hoe groot het verschil is in lengte tussen jongens en meisjes.

Opgave 23

Gegeven is onderstaande cumulatieve frequentiepolygoon voor het aantal uren televisiekijken per week door jongens respectievelijk meisjes.



Onderzoek hoe groot het verschil is tussen jongens en meisjes in het aantal uren televisiekijken.

§ 4.5.2.5 Om te onthouden

Om te bepalen of er sprake is van een (groot) verschil tussen twee groepen op een ordinale variabele bereken je het maximale cumulatieve percentageverschil. Met behulp van vuistregels geef je een oordeel over de omvang van het verschil tussen de twee groepen.

Maximale verschil in cumulatief percentage ($\max V_{cp}$)

- $\max V_{cp} > 40$:groot verschil.
- $20 < \max V_{cp} \leq 40$: middelmatig verschil.
- $\max V_{cp} \leq 20$: gering verschil.

§ 4.5.3 Op een kwantitatieve variabele met effectgrootte

§ 4.5.3.1 Introductie

Onderstaande tabel heeft betrekking op het gewicht bij jongeren (68 jongens en 84 meisjes) in kilogram.

	Jongen	**Meisje**
Gemiddelde	65,2	56,8
Standaardafwijking	9,24	6,64

§ 4.5.3.2 Centrale vraag

Hoe kun je aan de hand van de tabel bepalen of er sprake is van een (groot) verschil tussen jongens en meisjes voor wat betreft het gewicht?

§ 4.5.3.3 Antwoord op de centrale vraag

Om de centrale vraag te beantwoorden, bereken je de effectgrootte.
Dat doe je met de volgende formule:

effectgrootte $E = \frac{\overline{X}_1 - \overline{X}_2}{\frac{1}{2}(S_1 + S_2)}$

met $\overline{X}_1$ en $\overline{X}_2$ de steekproefgemiddelden ($\overline{X}_1 \geq \overline{X}_2$)

en S_1 en S_2 de steekproefstandaardafwijkingen.

In het voorbeeld over het gewicht van jongeren (jongens en meisjes) geldt dat het verschil in gemiddelden gelijk is aan 8,4 kilogram.
Voor de jongens is de standaardafwijking gelijk aan 9,24 kilogram. Voor de meisjes is dit 6,64 kilogram. Het gemiddelde van deze twee waarden is 7,94 kilogram.
De effectgrootte is dus gelijk aan 8,4 / 7,94 = 1,06.

Om op basis van deze waarde een oordeel te geven over de omvang van het verschil tussen beide groepen, maak je weer gebruik van vuistregels. Deze vuistregels zijn:

- $E > 0{,}8$: groot verschil.
- $0{,}4 < E \leq 0{,}8$: middelmatig verschil.
- $E \leq 0{,}4$: gering verschil.

§ 4.5.3.4 Oefenen

Opgave 24
We willen nagaan hoe groot het verschil is in de IQ-scores van volwassenen die een hbo-opleiding hebben afgerond en de IQ-scores van volwassenen met een mbo-opleiding.
Van deze twee groepen zijn onderstaande gegevens bekend.

	Hbo	**Mbo**
Gemiddelde	122	108
Standaardafwijking	10	15

Onderzoek met behulp van de effectgrootte hoe groot het verschil is in IQ-scores van deze twee groepen.

Opgave 25
In onderstaande tabel staan voor twee restaurants het gemiddelde fooibedrag en de standaardafwijking.

	Restaurant A	**Restaurant B**
Gemiddelde	10	16
Standaardafwijking	5	10

Onderzoek met behulp van de gegevens in de tabel hoe groot het verschil is in het fooibedrag tussen beide restaurants.

§ 4.5.3.5 Om te onthouden

Om na te gaan of er sprake is van een (groot, middelmatig of gering) verschil tussen twee groepen op een kwantitatieve variabele kun je de effectgrootte gebruiken.
Met behulp van vuistregels geef je een oordeel over de omvang van het verschil tussen de twee groepen.

Effectgrootte $E = \dfrac{\overline{X}_1 - \overline{X}_2}{\frac{1}{2}(S_1 + S_2)}$

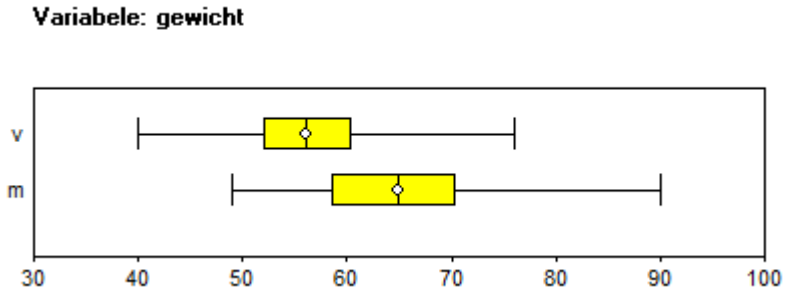
met $\overline{X}_1$ en $\overline{X}_2$ de steekproefgemiddelden ($\overline{X}_1 \geq \overline{X}_2$) en
S_1 en S_2 de steekproefstandaardafwijkingen.

- $E > 0{,}8$: groot verschil.
- $0{,}4 < E \leq 0{,}8$: middelmatig verschil.
- $E \leq 0{,}4$: gering verschil.

§ 4.5.4 Op een kwantitatieve variabele (vergelijken van boxplots)

§ 4.5.4.1 Introductie

Soms heb je niet alle benodigde informatie om de effectgrootte te kunnen berekenen, maar beschik je wel over de boxplot per groep (of kun je die maken).
In deze situatie kun je de twee boxplots met elkaar vergelijken.

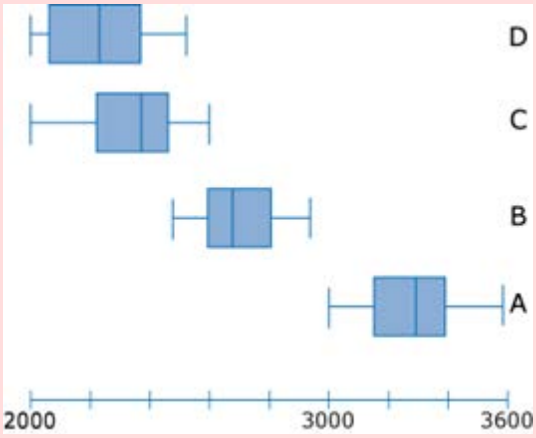


§ 4.5.4.2 Centrale vraag

Hoe kun je aan de hand van de boxplot bepalen of er sprake is van een groot verschil tussen jongens en meisjes voor wat betreft het gewicht?

§ 4.5.4.3 Antwoord op de centrale vraag

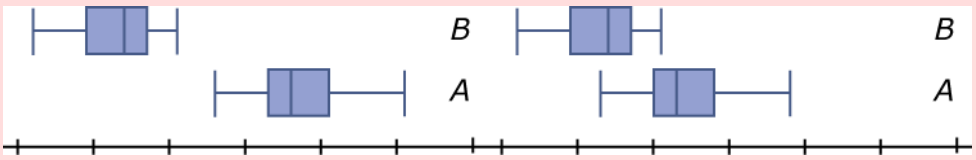
We nemen eerst een aanloopje om de centrale vraag te beantwoorden.
Hieronder zie je boxplots die het aantal branduren van vier type lampen beschrijven:



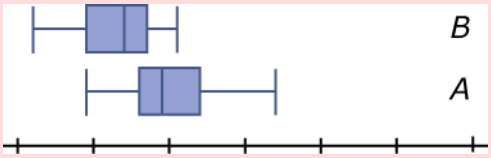
Je ziet onmiddellijk dat lampen van het type A een langere brandtijd hebben dan die van alle andere types. Immers zelfs de laagste gemeten brandtijd van een lamp van dit type is langer dan elke hoogste gemeten brandtijd van de andere types.
Maar hoe zit het als je de types B en C vergelijkt? Van die types overlappen de boxplots elkaar gedeeltelijk. Maar je ziet ook dat 75 procent van de lampen van type B een langere brandtijd heeft dan alle lampen van type C. De conclusie dat de lampen van type B meestal langer meegaan dan die van type C is wel gerechtvaardigd.
Bij het vergelijken van de types C en D is het trekken van een gerechtvaardigde conclusie veel moeilijker. De overlap van beide boxplots is zo groot, dat de brandtijd van alle lampen van type D valt binnen de boxplot van type C. Wel kun je zeggen dat de 50 procent lampen van type C die het langst meegaan een langere brandtijd hebben dan de 75 procent kortst brandende lampen van type D.

Uit het voorbeeld hierboven blijkt dat we bij het vergelijken van boxplots vaak kijken naar de overlap. Hieronder zie je drie situaties getekend waarin je een uitspraak kunt doen.

Situatie 1:
De boxen van A en B overlappen elkaar niet: het verschil is groot.



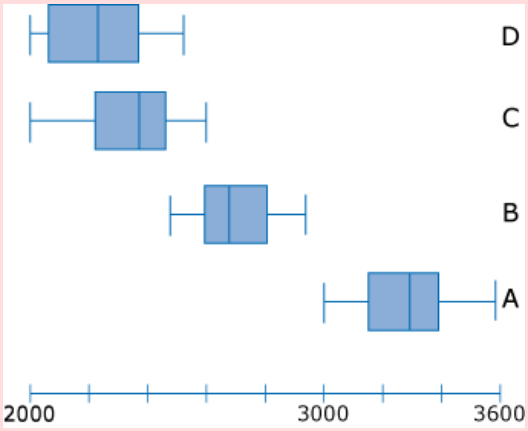
Situatie 2:
De boxen van A en van B overlappen elkaar en een mediaan van een boxplot ligt buiten de box van de andere boxplot: het verschil is middelmatig.



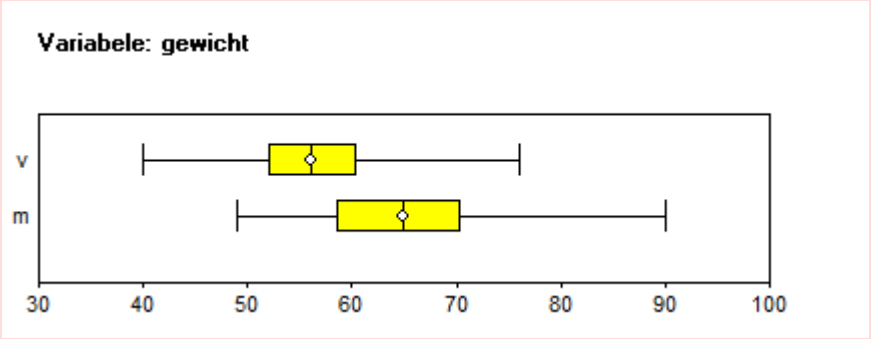
Situatie 3:
De boxen overlappen en voor beide medianen geldt dat deze binnen de box van de ander ligt: het verschil is gering.

Terug naar het voorbeeld van de lampen.
Volgens de vuistregels zoals geformuleerd in de drie situaties, zeggen we nu:

- Het verschil tussen C en D is gering, want de mediaan van D ligt in de box van C (en omgekeerd ligt die van C niet buiten de box van D).
- Het verschil tussen B en C is groot.



Kunnen we op basis van onderstaande boxplots en de vuistregels nu zeggen dat er groot verschil is tussen jongens en meisjes voor wat betreft het gewicht?

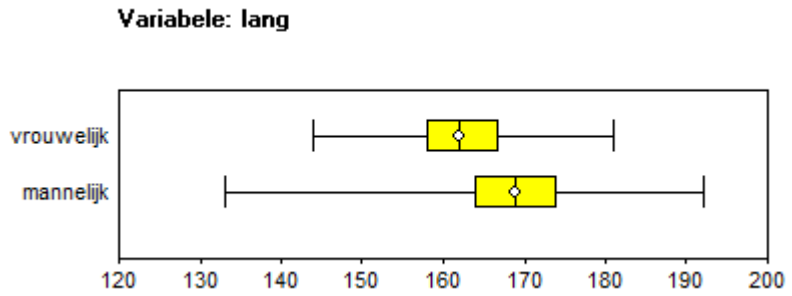


Samengevat is het antwoord deze vraag: de boxen overlappen, maar de medianen liggen niet in de andere boxen, dus er is sprake van een middelmatig verschil.
Zoals je ziet is het mogelijk dat het verschil bij de effectgrootte groot, terwijl het bij het vergelijken van boxplots middelmatig is. In zo'n situatie kun je de conclusie het beste baseren op de effectgrootte. Deze gebruikt namelijk meer informatie uit de steekproef.

§ 4.5.4.4 Oefenen

Opgave 26

Van een aselecte steekproef van 404 kinderen is het geslacht bekend en hun lengte in centimeters (variabele: *lang*). Hieronder zie je de boxplot voor de lengte gesplitst op geslacht.



Bepaal door het vergelijken van de boxplot hoe groot het verschil is tussen jongens en meisjes voor wat betreft hun lengte.

Opgave 27

Van een aselecte steekproef van 404 kinderen is het geslacht (variabele: *sekse*) bekend en hun schoenmaat (in Engelse maat). Hieronder zie je de kentallen voor de schoenmaat gesplitst op geslacht.

Sekse	**Mannelijk**	**Vrouwelijk**
Aantal waarnemingen	209	195
Gemiddelde	8,07	5,39
Mediaan	8,0	5,5
Modus	8,0	6,0
Minimum	3,0	1,0
Maximum	12,0	9,0
SDn-1	1,489	1,233
SDn	1,485	1,230
VARn-1	2,217	1,521
VARn	2,206	1,513
Eerste kwartiel	7,00	4,50
Derde kwartiel	9,00	6,00
Kwartielafstand	2,00	1,50

Bepaal door het vergelijken van de boxplot hoe groot het verschil is tussen jongens en meisjes voor wat betreft hun schoenmaat.

§ 4.5.4.5 Om te onthouden

Om na te gaan of er sprake is van een (groot, middelmatig of gering) verschil tussen twee groepen op een kwantitatieve variabele kun je boxplots vergelijken. Met behulp van vuistregels geef je oordeel over de omvang van het verschil tussen de twee groepen.

Twee boxplots vergelijken

- Als de boxen elkaar niet overlappen, dan is het verschil groot.
- Als de boxen elkaar wel overlappen en de mediaan van de ene boxplot buiten de box van de andere boxplot ligt, dan is het verschil middelmatig.
- In alle andere gevallen is het verschil gering.

§ 4.5.4.6 Geïntegreerd oefenen

Opgave 28

Geslacht	**Jongen**	**Meisje**
Aantal waarnemingen	24472	25599
Gemiddelde	14,8	13,7
Mediaan	12,0	11
Modus	10	10
Minimum	0	0
Maximum	70	70
SDn-1	10,60	10,24
SDn	10,60	10,24
VARn-1	112,42	104,89
VARn	112,42	104,89
Eerste kwartiel	7,0	7,0
Derde kwartiel	20,0	19,0
Kwartielafstand	13,0	12,0

Onderzoek met behulp van effectgrootte en ook met het vergelijken van boxplots hoe groot het verschil is tussen jongens en meisjes wat betreft het aantal uren televisiekijken per week.

Opgave 29

Twee lesgroepen doen examen. Voor het examen kun je 90 punten behalen.
Als je 45 of meer punten hebt, krijg je een voldoende.
Van de uitslag is het volgende bekend.

	Lesgroep 1	Lesgroep 2
Onvoldoende	8	3
Voldoende	20	19

In de tabel kun je bijvoorbeeld aflezen dat er in lesgroep 1 acht leerlingen een onvoldoende hebben gehaald voor het examen.

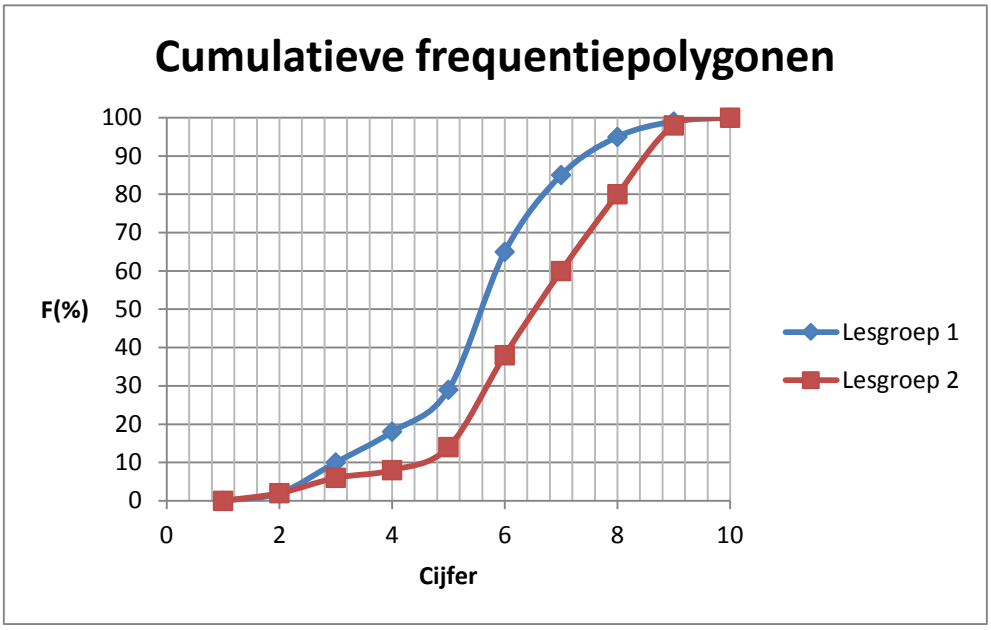
a. Kun je aan de hand van deze tabel aangeven of er sprake is van groot verschil tussen beide lesgroepen?

Van beide groepen zijn de resultaten op weergegeven in onderstaand cumulatief frequentiepolygoon.

b. Laat zien dat deze polygonen in overeenstemming zijn met bovenstaande kruistabel. Motiveer je antwoord.

c. In welke lesgroep is het examen het beste gemaakt? Motiveer je antwoord.

d. Bereken het maximale cumulatieve percentageverschil ($\max V_{cp}$) en bepaal of er sprake is van een groot verschil.

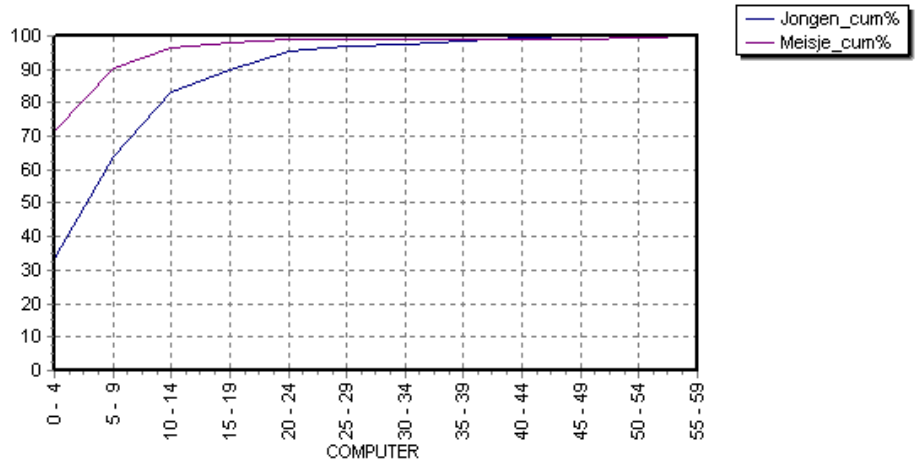


Opgave 30

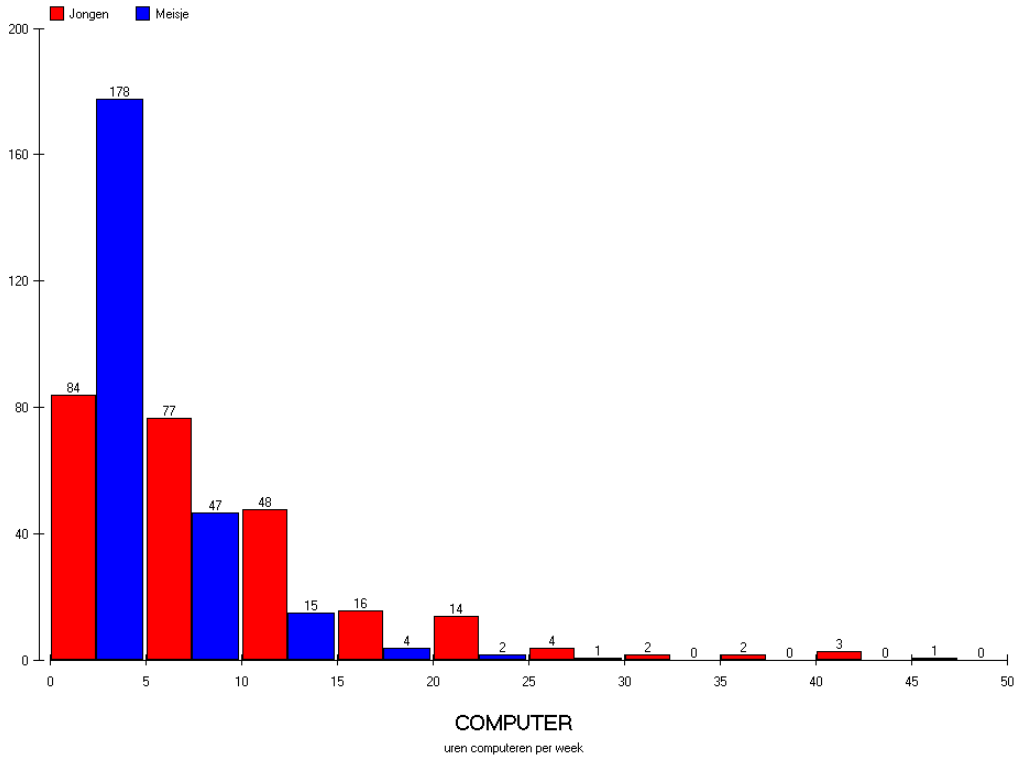
Ruim 50000 leerlingen uit leerjaar 1 en 2 van het voortgezet onderwijs vullen een CBS-enquête in.
Een van de vragen is ‘hoeveel uur per week zit je achter de computer?’
Uit al deze gegevens wordt een steekproef genomen van 500 leerlingen. Op de volgende bladzijde zie je gegevens van deze 500 leerlingen, met 251 jongens en 249 meisjes.
We geven je drie bronnen met informatie over het computergebruik. Steeds zie je gegevens over jongens en meisjes.

Figuur 1: cumulatief frequentiepolygoon

De linker grafiek betreft de meisjes, de rechter grafiek betreft de jongens.
Bij een cumulatief frequentiepolygoon moet je eigenlijk horizontaal steeds de rechtergrens lezen; de getallen die bij de verticale stippellijntjes horen zijn dus achtereenvolgens: 4, 9, 14, 19, .., 64.)



Figuur 2: geclusterd staafdiagram



Figuur 3: tabel met centrummaten

Geslacht	Jongen	Meisje
Aantal waarnemingen	251	249
Gemiddelde	8,7	3,9
Mediaan	7	2
Modus	5	0
Minimum	0	0
Maximum	45	57
SD	7,87	6,13
Eerste kwartiel	3,0	1,0
Derde kwartiel	12,0	5,0

We kijken naar het verschil tussen jongens en meisjes.
Op basis van de figuren 1 tot en met 3 kun je beslissen dat jongens meer computeren dan meisjes.

a. Geef een duidelijke uitleg hoe je dit in figuur 1 afleest. Doe dat ook voor figuur 2 en 3.

Of het verschil tussen jongens en meisjes groot is of klein kun je berekenen.

b. Bereken $\max V_{cp}$ en bepaal of er sprake is van groot verschil tussen jongens en meisjes.

c. Bereken de effectgrootte en bepaal of er volgens de effectgrootte sprake is van groot verschil is tussen jongens en meisjes.

Opgave 31

Bij een onderzoek naar de gevolgen van het pilgebruik wordt onder andere de bloeddruk van vrouwen gemeten. Hieronder zie je gegevens (uitgedrukt in procenten) over de bloeddruk van vrouwen (in de leeftijdscategorie 25-34 jaar) die al meer dan een jaar de pil slikken (*users*) en vrouwen die dat niet doen (*non-users*).

Bloeddruk (in mm)	**Non-user**	**User**
Onder 90	1	
90-95	1	
95-100	5	4
100-105	11	5
105-110	11	10
110-115	17	15
115-120	18	17
120-125	11	15
125-130	9	12
130-135	7	10
135-140	4	5
140-145	2	4
145-150	2	2
150-155	1	1
Totaal (in procenten)	100	100

Onderzoek hoe groot het verschil in bloeddruk is tussen de users en de non-users.
Kies hiervoor een geschikte maat.

§ 4.6 Samenhang tussen twee kwantitatieve variabelen

- Ga naar www.youtube.com.
- Zoek op ‘Beschrijvende statistiek -correlatie’.
- Bekijk het filmpje hierover van de WiskundeAcademie.

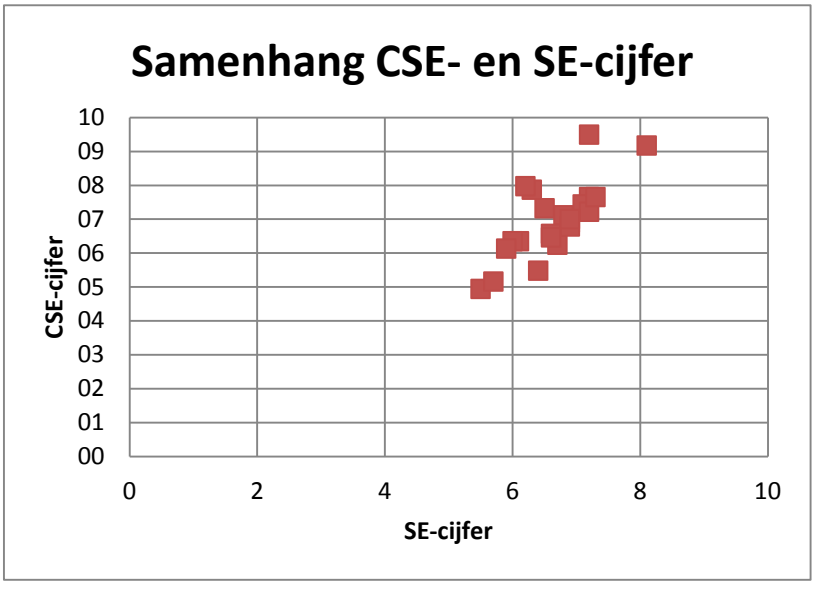
Deze paragraaf gaat over de samenhang tussen twee kwantitatieve variabelen.
Deze samenhang kun je bestuderen door eerst te kijken naar de bijbehorende puntenwolk.
De mate van samenhang druk je uit in een getal, de zogenaamde correlatiecoëfficiënt. Deze maat voor samenhang komt niet voor in de eindtermen van het Centraal Schriftelijk Examen (CSE).
Daarom presenteren we dit deel (paragraaf 4.6.1) als extra stof.
De zogenaamde trendlijn behoort wel tot de examenstof.
Uitleg over een trendlijn staat in paragraaf 4.6.2.

EXTRA

§ 4.6.1 Correlatiecoëfficiënt

§ 4.6.1.1 Introductie

Een docent bekijkt voor een examengroep de samenhang tussen de schoolexamencijfers van de leerlingen en hun cijfers voor het centraal eindexamen. De puntenwolk staat hieronder afgebeeld.
In deze context is het logisch om het SE-cijfer op de x-as te plaatsen en het CSE-cijfer op de y-as, omdat het CSE wordt afgenomen na het SE (schoolexamen).



De correlatiecoëfficiënt is 0,75. Je hoeft de waarde van de correlatiecoëfficiënt niet zelf te kunnen berekenen, daarvoor gebruik je ICT als hulpmiddel.

§ 4.6.1.2 Centrale vraag

Wat zegt de correlatiecoëfficiënt van 0,75 over de samenhang tussen de SE-cijfers en de CSE-cijfers?

De correlatiecoëfficiënt tussen twee kwantitatieve variabelen heeft altijd een waarde tussen -1 en +1.

- Liggen alle punten van de puntenwolk precies op een stijgende lijn, dan is de correlatiecoëfficiënt gelijk aan 1. Er is een perfecte positieve lineaire samenhang tussen de twee variabelen.
- Liggen alle punten van de puntenwolk precies op een dalende lijn, dan is de correlatiecoëfficiënt -1. Er is een perfecte negatieve lineaire samenhang tussen de twee variabelen.
- Is er nauwelijks een lineair verband tussen de twee variabelen, dan is de correlatiecoëfficiënt ongeveer 0.

Om aan de waarde van de correlatiecoëfficiënt ( ) een oordeel toe te kennen, maak je weer gebruik van vuistregels:

- $r \leq -0{,}7$: sterke negatieve samenhang.
- $-0{,}7 < r \leq -0{,}3$: matige negatieve samenhang.
- $-0{,}3 < r < 0$: zwakke negatieve samenhang.

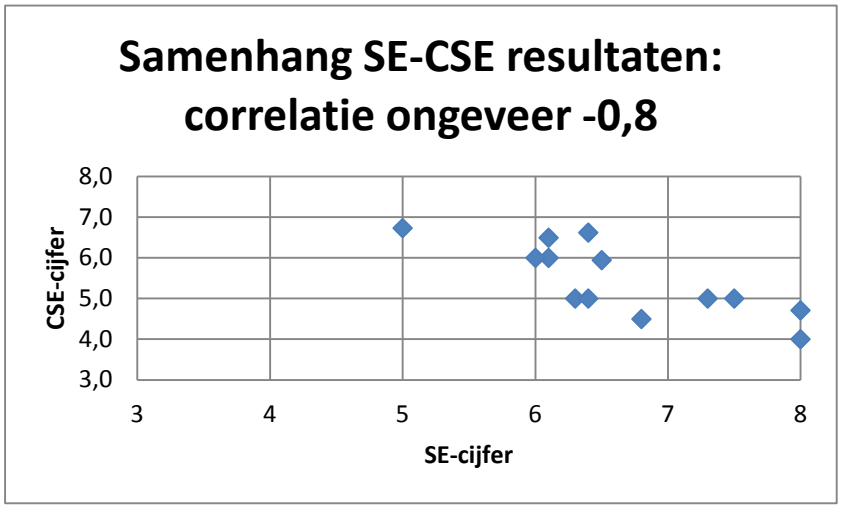
- $0 < r < 0{,}3$: zwakke positieve samenhang.
- $0{,}3 < r < 0{,}7$: matige positieve samenhang.
- $r \geq 0{,}7$: sterke positieve samenhang.

De waarde van 0,75 duidt op een sterke positieve samenhang tussen het SE-cijfer en het CSE-cijfer van de betreffende leerlingen.

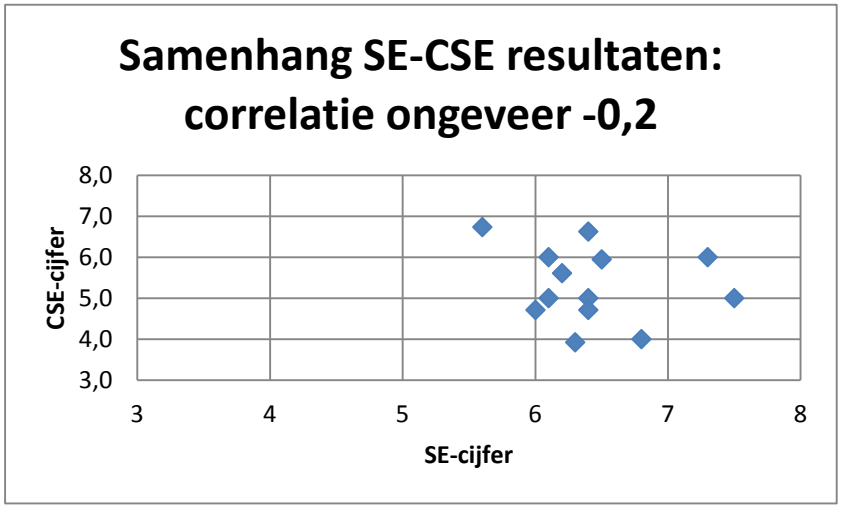
§ 4.6.1.3 Antwoord op de centrale vraag

Om enig gevoel te krijgen bij de waarde van een correlatiecoëfficiënt staan hierna vijf puntenwolken afgebeeld, met ieder een eigen waarde van de correlatiecoëfficiënt.

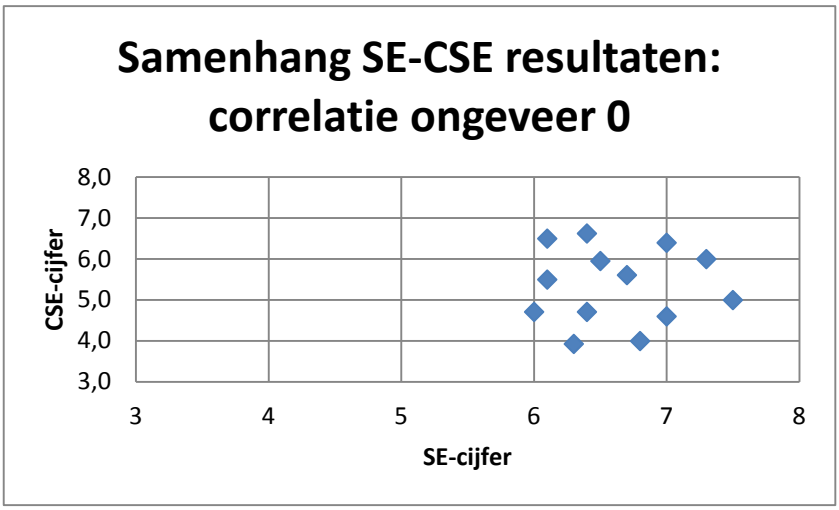
In onderstaande figuur geldt in het algemeen dat hoe hoger het SE-cijfer, hoe lager het CSE-cijfer; er bestaat dus een negatieve samenhang.
Als je door de puntenwolk een zo goed mogelijk passende recht lijn tekent, dan liggen de punten in het algemeen niet erg ver van die lijn; de mate van lineaire samenhang is dus sterk.
De correlatiecoëfficiënt is -0,8 (sterke negatieve lineaire samenhang).



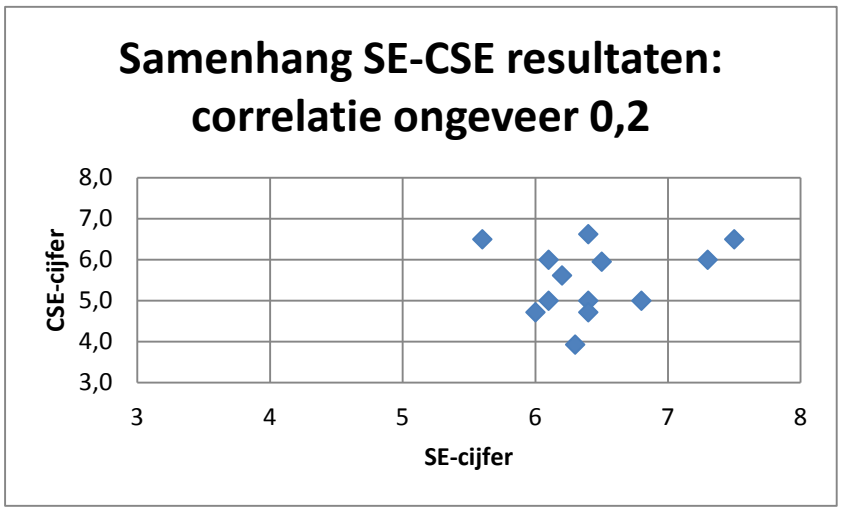
In onderstaande figuur geldt in het algemeen dat hoe hoger het SE-cijfer, hoe lager het CSE-cijfer; er bestaat dus een negatieve samenhang.
Als je door de puntenwolk een zo goed mogelijk passende rechte lijn tekent, dan liggen de punten in het algemeen behoorlijk ver van die lijn; de mate van lineaire samenhang is dus zwak.
De correlatiecoëfficiënt is -0,2 (zwakke negatieve lineaire samenhang).



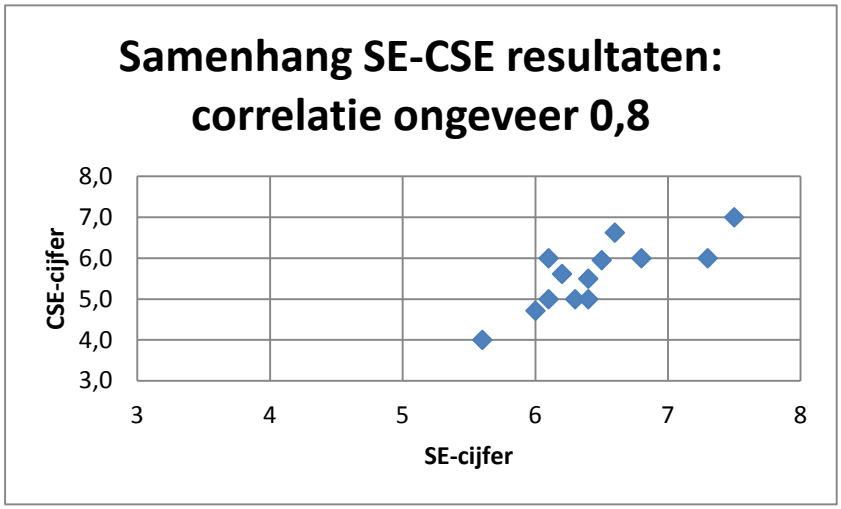
In onderstaande figuur bestaat er geen lineaire samenhang tussen het SE-cijfer en het CSE-cijfer; de correlatiecoëfficiënt is 0.
Elke lijn die je door de puntenwolk kunt tekenen past even goed bij de puntenwolk.
In deze situatie heeft het behaalde SE-cijfer geen enkele voorspellende waarde voor het CSE-cijfer.



In onderstaande figuur geldt in het algemeen dat hoe hoger het SE-cijfer, hoe hoger het CSE-cijfer; er bestaat dus een positieve samenhang.
Als je door de puntenwolk een zo goed mogelijk passende rechte lijn tekent, dan liggen de punten in het algemeen behoorlijk ver van die lijn; de mate van lineaire samenhang is dus zwak.
De correlatiecoëfficiënt is 0,2 (een zwakke positieve lineaire samenhang).



In onderstaande figuur geldt in het algemeen dat hoe hoger het SE-cijfer, hoe hoger het CSE-cijfer; er bestaat dus een positieve samenhang.
Als je door de puntenwolk een zo goed mogelijk passende rechte lijn tekent, dan liggen de punten in het algemeen niet erg ver van die lijn; de mate van lineaire samenhang is dus sterk.
De correlatiecoëfficiënt is 0,8 (een sterke positieve lineaire samenhang).



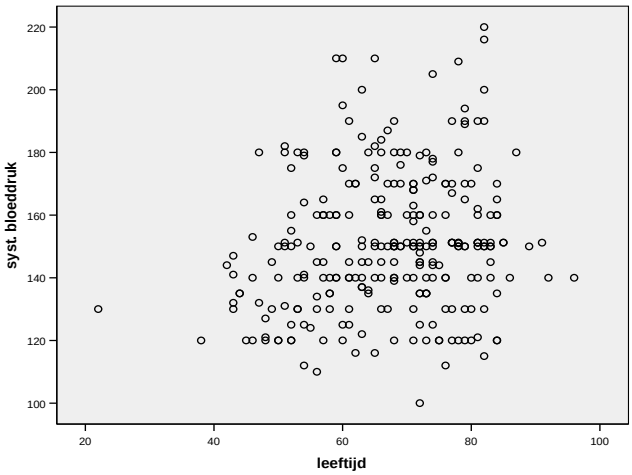
Opmerkingen

- Een correlatiecoëfficiënt (in de buurt) van 0 betekent niet dat er geen samenhang bestaat tussen de twee variabelen. De enige conclusie die je kunt trekken is dat er geen s lineaire samenhang is tussen de twee variabelen. Er kan dus wel sprake zijn van een andere vorm van samenhang.
- Een correlatiecoëfficiënt zegt uitsluitend iets over de samenhang tussen twee variabelen en niets over een eventuele causale relatie (=oorzaak-gevolgrelatie) tussen de twee variabelen.

§ 4.6.1.4 Oefenen

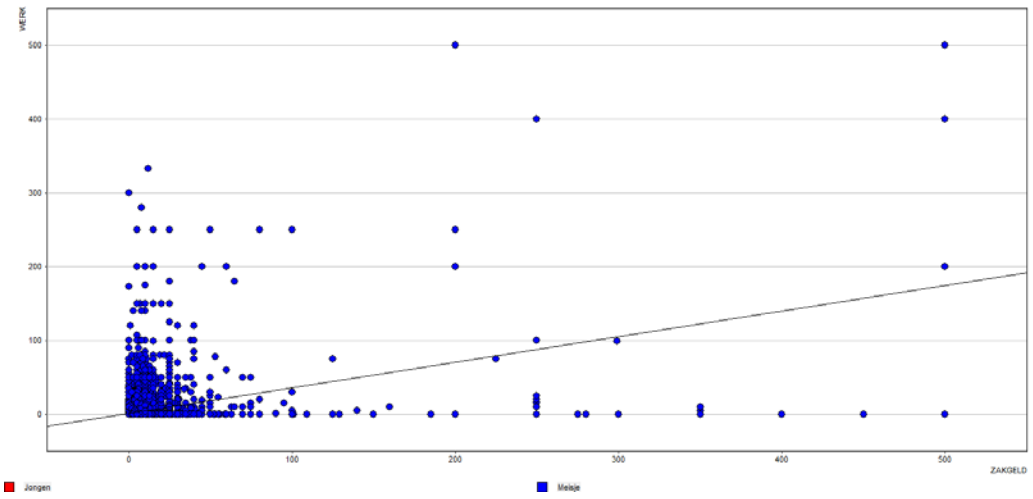
Opgave 32

Hieronder zie je een puntenwolk waarin ieder punt de leeftijd en de bloeddruk van een persoon geeft.
Geef een schatting van de correlatiecoëfficiënt.



Opgave 33

Voor meisjes in leerjaar 1 is de onderstaande puntenwolk opgesteld.



De correlatiecoëfficiënt is 0,38.
Welke conclusie kun je trekken over de mate van samenhang tussen deze twee variabelen?

§ 4.6.1.5 Om te onthouden

Om een uitspraak te doen over de mate van samenhang tussen twee kwantitatieve variabelen bereken je de correlatiecoëfficiënt (r).

Om aan de waarde van de correlatiecoëfficiënt een oordeel toe te kennen, maak je weer gebruik van vuistregels:

- $r \leq -0{,}7$: sterke negatieve samenhang.
- $-0{,}7 < r \leq -0{,}3$: matige negatieve samenhang.
- $-0{,}3 < r < 0$: zwakke negatieve samenhang.

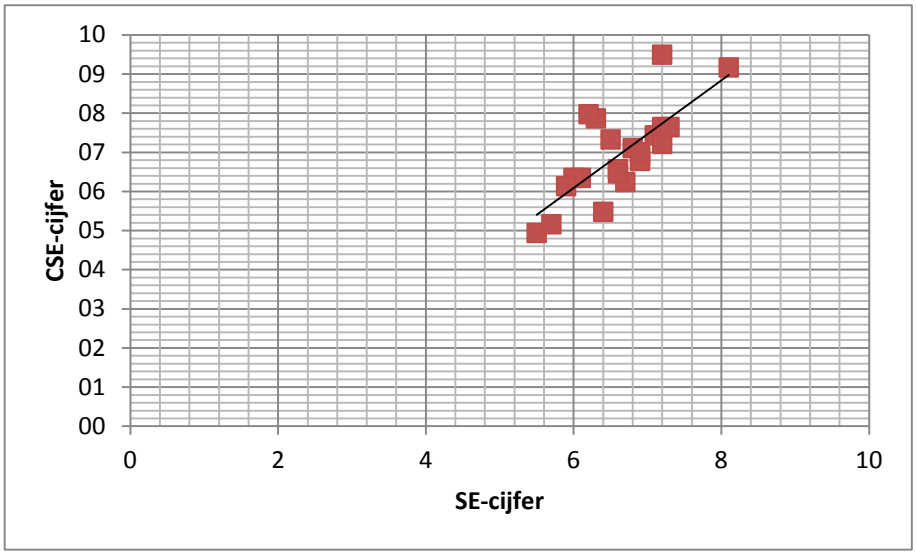
- $0 < r < 0{,}3$: zwakke positieve samenhang.
- $0{,}3 < r < 0{,}7$: matige positieve samenhang.
- $r \geq 0{,}7$: sterke positieve samenhang.

EINDE EXTRA

§ 4.6.2 Trendlijn

§ 4.6.2.1 Introductie

In onderstaande figuur staat ook de trendlijn afgebeeld.
Dit is de lijn die het beste past bij de puntenwolk.



Voor de trendlijn geldt de formule: CSE-cijfer = 1,04 * SE-cijfer - 0,31
Je hoeft deze formule niet zelf te berekenen op basis van alle gegevens;
daarvoor gebruik je ICT als hulpmiddel.

§ 4.6.2.2 Centrale vraag

Wat zegt de trendlijn in een puntenwolk over de samenhang tussen de twee variabelen?

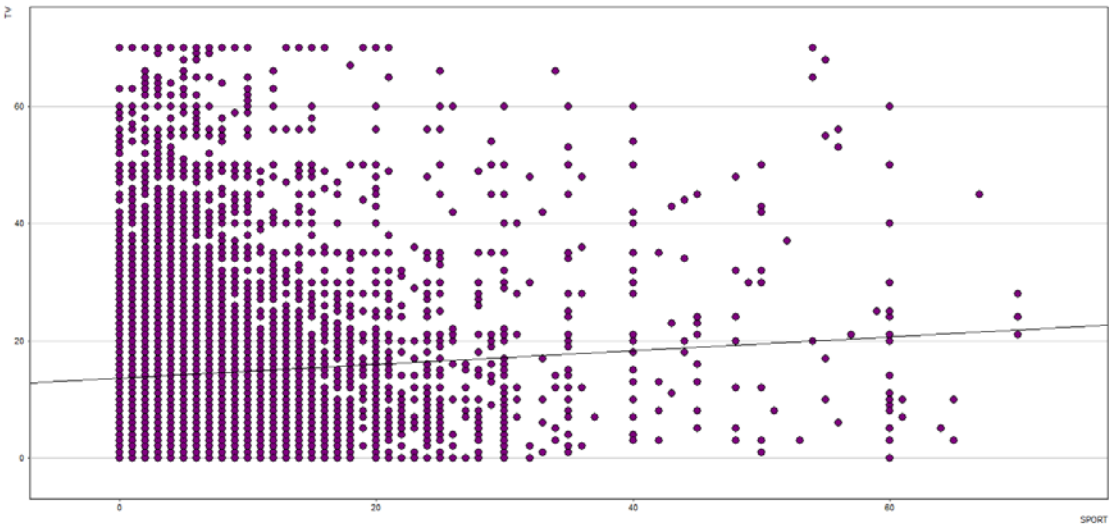
§ 4.6.2.3 Antwoord op de centrale vraag

Met behulp van deze formule kun (zo goed mogelijk) voorspellen welk CSE-cijfer een leerling gaat halen wanneer je zijn of haar SE-cijfer kent.
Bijvoorbeeld: als een leerling als SE-cijfer 6,2 heeft, dan is de beste voorspelling voor het CSE-cijfer dat deze leerling gaat halen gelijk aan 6,1 (= 1,04 * 6,2 - 0,31).

§ 4.6.2.4 Oefenen

Opgave 34

Gegeven is onderstaande puntenwolk met op de x-as het aantal uren sport kijken per week en op de y-as het aantal uren televisiekijken per week.



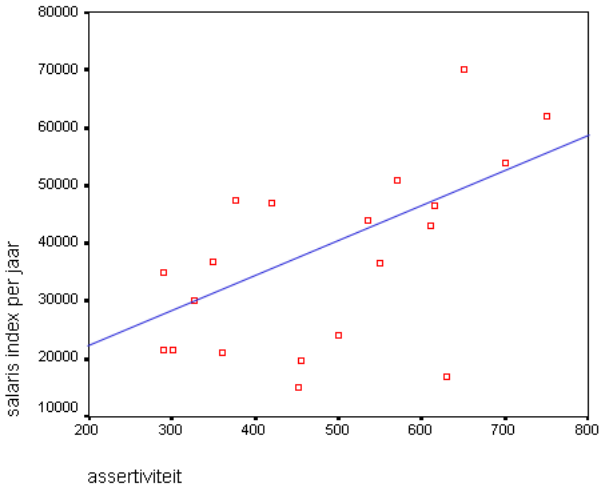
De formule voor de trendlijn is: TV = 13,71 + 0,12 * SPORT

Bereken TV als:

a. SPORT = 5.
b. SPORT = 10.
c. SPORT = 20.

Opgave 35

Twee collega's hebben een discussie met elkaar: de ene is van mening dat je meer zult verdienen als je assertief bent en de ander denkt dat je op sommige ogenblikken juist beter je mond kunt houden. Ze besluiten daarom de proef op de som te nemen en doen in hun bedrijf een kleine analyse over het loon en het assertiviteitsgehalte van hun collega's. Ze meten de assertiviteit via een vragenlijst die ze op internet hebben gevonden.

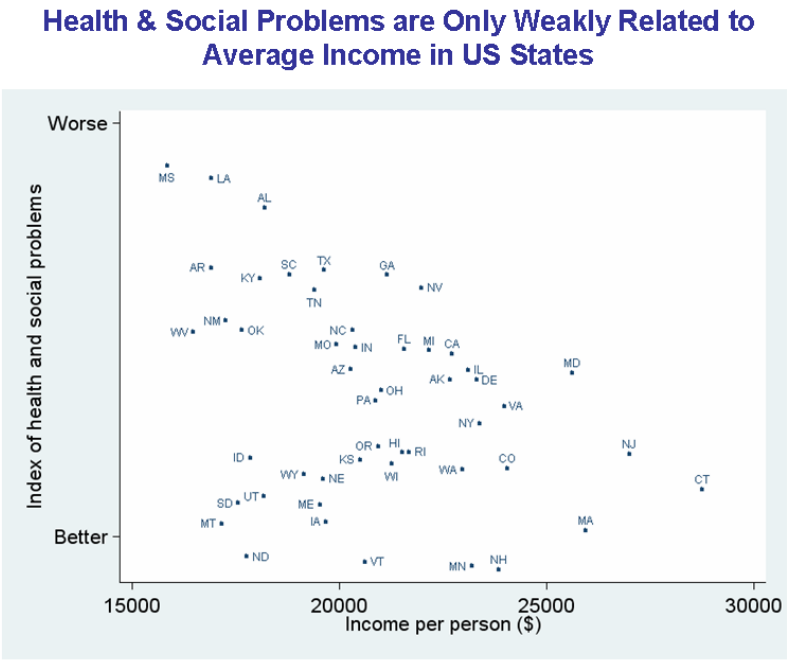


De formule voor de trendlijn is: S = 60·A + 10000.
Hierin is S de salarisindex per jaar en A de assertiviteit.

a. Bereken de verwachte salarisindex voor iemand met een assertiviteit van 500.
b. Bereken de verwachte assertiviteit van iemand met een salarisindex van 50000.

Opgave 36

Hieronder zie je een figuur waarin per staat binnen de Verenigde Staten het inkomen per persoon is uitgezet tegen een index voor gezondheids- en sociale problemen.



Source: Wilkinson & Pickett, The Spirit Level (2009)

www.equalitytrust.org.uk ™ Equality Trust

Welke conclusie kun je trekken ten aanzien van de samenhang tussen inkomen enerzijds en gezondheids- en sociale problemen anderzijds?

Opgave 37

- Ga naar www.youtube.com.
- Zoek en bekijk het filmpje ‘The danger of mixing up causality and correlation’.
- Leg uit wat de belangrijkste boodschap is van dit filmpje.

§ 4.6.2.5 Om te onthouden

Als je de samenhang tussen twee kwantitatieve variabelen wilt uitdrukken in een formule, bereken je de formule van de trendlijn. Dit is de lijn die het beste past bij de puntenwolk van de twee variabelen. Met behulp van de formule van de trendlijn kun je uitrekenen welke waarde een variabele naar verwachting heeft als je de waarde van de andere variabele kent.

§ 4.7 Gemengde opdrachten

Opgave 38

Zo'n 15 procent van de jongeren is te zwaar.
Maar als je het jongeren zelf vraagt, vinden ze bijna allemaal dat ze in goede gezondheid verkeren.
Dat blijkt uit cijfers van het CBS over 2010 tot en met 2012.

Veronderstel dat in een aselecte steekproef onder 500 jongeren in Nederland is gevraagd hoe zij hun gezondheid ervaren. De mogelijke antwoorden zijn: '*slecht*', '*matig*', '*voldoende*' en '*goed*'.
462 jongeren beantwoorden de vraag met '*goed*'.

a. Wat is de populatie?
b. Wat is het meetniveau van de variabele *ervaren gezondheid*?
c. Bereken het 95%-betrouwbaarheidsinterval voor de proportie jongeren in de populatie dat hun gezondheid als goed ervaart.
d. Is het 99%-betrouwbaarheidsinterval breder of juist minder breed dan het 95%-betrouwbaarheidsinterval?

Opgave 39

Wie een facelift laat uitvoeren, ziet er na de behandeling gemiddeld 3 jaar jonger uit. Wie meer jaren wil liegen, moet meer operaties laten uitvoeren en dus ook meer geld uitgeven. Dat is de conclusie van Amerikaanse en Canadese specialisten na onderzoek waarin aan 50 onafhankelijke proefpersonen werd gevraagd om foto's van voor en na de operatie van 49 patiënten te beoordelen.
De patiënten waren tussen de 42 en 73 jaar oud op het moment van de operatie, met een gemiddelde leeftijd van 57 jaar. De patiënten werden ongeveer 2,1 jaar jonger ingeschat op de foto's voor de operatie en 5,2 jaar jonger op de foto's na de operatie. Gemiddeld leken de patiënten dus 3,1 jaar jonger na de operatie dan ervoor.

a. Wat is de populatie?
b. Hoe groot is de steekproefomvang?

Veronderstel dat de standaardafwijking van het gemiddeld aantal jaar dat men er na de faceliftoperatie jonger uitziet gelijk is aan 0,8.

c. Bereken het 95%-betrouwbaarheidsinterval voor het gemiddelde aantal jaren dat de patiënten er na een faceliftoperatie jonger uit zien.
d. Bereken hoe groot de steekproefomvang zou moeten zijn om met 95 procent betrouwbaarheid het gemiddeld aantal jaren dat men er jonger uitziet na een faceliftoperatie op 1 decimaal nauwkeurig te kunnen berekenen.

Opgave 40

Er is veel onderzoek gedaan naar barbecueën, zij het bijna altijd door producenten die zelf baat hebben bij een bepaalde uitkomst. Toch zijn de resultaten interessant.
Uit recent onderzoek van Intomart in opdracht van producent Campingaz blijkt dat 44 procent van de ondervraagde groep Nederlanders en Belgen 1 tot 6 keer per jaar barbecuet. Een kwart van de ondervraagde doet dit 6 tot 11 keer per jaar.
Veronderstel dat de volgende gegevens zijn verzameld:

Aantal keren barbecueën	Nederlanders	Belgen
0	25	18
1-5	60	36
6-10	30	25
11-20	15	12
Totaal	130	90

a. Bereken het maximale cumulatieve percentageverschil: $\max V_{cp}$.
b. Teken de boxplots en vergelijk deze met behulp van de vuistregels op het formuleblad.
c. Bereken de effectgrootte.
d. Welke conclusie kun je trekken met betrekking tot het verschil tussen beide groepen?

Opgave 41

Leasemaatschappij Business Lease Nederland weet het zeker: vrouwen zijn officieel slechtere chauffeurs dan mannen. Ze krijgen in elk geval veel vaker verkeersboetes dan mannen, blijkt uit een analyse van 11000 bekeuringen die het afgelopen half jaar zijn uitgedeeld. Vrouwen rijden vaker dan mannen door rood licht. "*Misschien dat vrouwen iets gehaaster rijden en van alles tegelijk doen. En dan letten ze misschien wat minder goed op*," aldus de directeur van het leasebedrijf.

Veronderstel dat Business Lease de onderstaande tabel heeft opgesteld.

	Niet bekeurd voor rijden door rood licht	Een of meerdere keren bekeurd voor rijden door rood licht
Vrouwen	48800	1200
Mannen	98000	2000

a. Bereken phi. Hoe groot is het verschil tussen vrouwen en mannen?

Waar vrouwen vaker door rood rijden en foutparkeren, rijden mannen vaker te hard. De gemiddelde 'mannelijke' boete bedraagt 71,97 euro. Vrouwen doen het iets beter met 70,20 euro.
Veronderstel dat de standaardafwijking voor de 'mannelijke' boete 8,30 euro bedraagt en die voor 'vrouwelijke' boete 5,70.
b. Bereken de effectgrootte. Hoe groot is het verschil tussen mannen en vrouwen?

Opgave 42

Kinderen van migrantenouders op gemengde scholen scoren hogere wiskundecijfers dan hun leeftijdsgenoten op ‘blanke’ scholen. Dit blijkt uit een onderzoek naar de wiskundeprestaties van 13-jarigen op elf Rotterdamse scholen voor het voortgezet onderwijs.
Volgens de onderzoekers laat de uitkomst een positief effect van de multiculturele samenleving zien. “*Kinderen kennen de straattaal en begrijpen elkaar. In de klas is het makkelijker elkaar te helpen bij het oplossen van wiskundeformules*”, concludeert onderwijssocioloog Sjaak Braster van de Erasmus Universiteit Rotterdam.
Voor het onderzoek zijn alleen wiskundecijfers geanalyseerd. “*Hierbij speelt de afkomst het minst een rol en dat geeft dus een zuiverder beeld. Bij het presteren bij een taalvak als Nederlands of Engels speelt herkomst veel meer een rol.*”
In dit onderzoek spelen verschillende variabelen een rol: de leeftijd van de kinderen, de herkomst van de ouders, het type school dat de leerlingen bezoeken en de wiskundecijfers die de kinderen halen.

a. Benoem het meetniveau van elk van de vier bovengenoemde variabelen.

Veronderstel dat onderstaande tabel kan worden opgesteld:

	Gemengde scholen	**’Blanke’ scholen**
Gemiddelde wiskundecijfer	6,8	6,2
Standaardafwijking	1,2	1,0

b. Bereken de effectgrootte. Hoe groot is het verschil tussen de twee typen scholen?

Volgens de onderzoekers laat de uitkomst een positief effect zien van de multiculturele samenleving.
c. Noem twee andere mogelijke verklaringen voor de gevonden resultaten.

Opgave 43

Een op de zeven vrouwen tussen de 25 en de 35 jaar kampt met burn-outverschijnselen.
"*Ze voelen zich enorm verantwoordelijk voor alles wat ze doen.*"
Er wordt veel gespeculeerd over de oorzaken. Het groeiende aantal mogelijkheden – studeren, reizen, werken – zou keuzestress veroorzaken. De hoge werkloosheid en het toenemend aantal tijdelijke contracten zouden tot extra prestatiedruk leiden. En de eindeloze stroom succesverhalen op sociale media zou jonge mensen het gevoel geven dat hun eigen leven nooit goed genoeg is.

Veronderstel dat onderstaande tabel kan worden opgesteld voor een aselecte steekproef onder vrouwen en mannen tussen de 25 en de 35 jaar.

	Wel burn-outverschijnselen	**Geen burn-outverschijnselen**
Vrouwen	20	120
Mannen	12	128

a. Bereken phi. Hoe groot is het verschil tussen vrouwen en mannen?

Voor jonge vrouwen komt daar bij dat zij vaak bepaalde karaktereigenschappen hebben die hen kwetsbaarder maakt voor overbelasting. Ze kunnen beter multitasken dan mannen, maar ze voelen zich ook nog eens enorm verantwoordelijk voor alles wat ze doen. Of het nu gaat om werk, relatie of uiterlijk.

Veronderstel dat onderstaande tabel geldt voor een aselecte steekproef onder vrouwen en mannen tussen de 25 en de 35 jaar.

Verantwoordelijkheidsgevoel	**Vrouwen**	**Mannen**
Klein	15	40
Gemiddeld	25	50
Groot	60	30
Enorm	40	20

b. Bereken het maximale cumulatieve percentageverschil. Hoe groot is het verschil tussen mannen en vrouwen?

Veronderstel dat er een puntenwolk wordt gemaakt met op de x-as de variabele *verantwoordelijkheidsgevoel* en op de y-as *burn-outverschijnselen*. Er wordt een trendlijn gevonden met een positieve helling en geconcludeerd dat een groot verantwoordelijkheidsgevoel leidt tot een grotere kans op een burn out.

c. Geef kritiek op de gevolgde werkwijze en op de conclusie.

§ 4.8 Terugblik

In deze module heb je geleerd om statistische uitspraken te doen over:

- Populatieproportie of populatiegemiddelde en de betrouwbaarheid ervan.
- Omvang van het verschil tussen twee groepen.
- Samenhang tussen twee kwantitatieve variabelen.

Centrale vragen die de revue hebben gepasseerd:

- Hoe kun je op basis van een steekproef een uitspraak doen over een populatieproportie en de betrouwbaarheid ervan kwantificeren?

Deze vraag is behandeld in paragraaf 4.3.

- Hoe kun je op basis van een steekproef een uitspraak doen over een populatiegemiddelde en de betrouwbaarheid ervan kwantificeren?

Deze vraag is behandeld in paragraaf 4.4.

- Hoe bepalen we of er sprake is van een (groot) verschil tussen twee groepen op een nominale , ordinale en kwantitatieve variabele?

Deze vraag is behandeld in paragraaf 4.5.

- Wat zegt de correlatiecoëfficiënt van twee kwantitatieve variabelen over de samenhang tussen deze variabelen? (EXTRA)
- Wat zegt de trendlijn in een puntenwolk over de samenhang tussen de twee variabelen?

Deze vragen zijn behandeld in paragraaf 4.6.

§ 4.9 Lessenserie: Statistiek op een groot gegevensbestand

Een eerdere versie van deze lessenserie staat op de website van het Centraal Bureau voor de Statistiek (CBS). Ga naar www.cbs.nl; kies achtereenvolgens voor ‘Informatie voor - Onderwijs’, ‘meer Lesplannen’, ‘Lesplannen naar vak -Wiskunde’, ‘Groot databestand beroepsbevolking’.

Gegevensbestand Enquête BeroepsBevolking (EBB) 2011
Het doel van het CBS bij deze EBB is het verstrekken van informatie over de relatie tussen mens en arbeidsmarkt. Hierbij worden kenmerken van personen in verband gebracht met hun huidige dan wel toekomstige positie op de arbeidsmarkt. Naast persoons- en huishoudkenmerken worden onder meer gegevens verzameld over de arbeidsmarktpositie, het arbeidsverleden en het opleidingsniveau van de Nederlandse bevolking van 15 tot 65 jaar. Men stelt dus vragen over:

- Persoonlijke kenmerken: zoals leeftijd, geslacht, opleiding en herkomst.
- Werkomstandigheden: zoals of je werk hebt, hoeveel uur je werkt, of je meer of minder zou willen werken en wat voor werk je doet.

Het bestand bevat 76746 records en onderstaande variabelen.
(Je kunt de gebruikte codes in VuStat zichtbaar maken door te klikken op de knop ‘Labels’.)

1. Geslacht
1 = man; 2 = vrouw.

2. Leeftijdsgroep
Leeftijd wordt berekend naar aanleiding van geboortedatum en vervolgens ingedeeld in tienjaarsklassen: 1 = 15-24 jaar; 2 = 25-34 jaar; 3 = 35-44 jaar; 4 = 45-54 jaar; 5 = 55-64 jaar.

3. Herkomstgroepering
De herkomstgroepering wordt bepaald naar aanleiding van het geboorteland van beide ouders.
Als beide ouders in Nederland zijn geboren is de respondent autochtoon. Als een van beide ouders in het buitenland is geboren is de respondent allochtoon. Wanneer beide ouders in een ander land zijn geboren, kijkt het CBS naar het geboorteland van de moeder om de herkomstgroepering te bepalen. Voorbeeld: moeder is geboren in Turkije, vader in Duitsland, dan heeft de respondent Turkije als herkomstland.
10 = autochtonen; 20 = westerse allochtonen; 31 = Turken en Marokkanen;
33 = Antillianen/Arubanen en Surinamers; 35 = overig niet-westerse landen; 99 = onbekend.

4. Beroepsbevolking
Heeft de respondent voor 12 uur per week of meer werk? Dan valt hij onder de werkzame beroepsbevolking.
Wil of kan de respondent niet werken voor 12 uur of meer per week (bijvoorbeeld scholieren en huisvrouwen/-mannen)? Dan valt hij onder de niet-beroepsbevolking.
Wil de respondent voor 12 uur of meer werken en doet hij dat nu niet? En kan hij op korte termijn beginnen en zoekt hij actief naar werk? Dan valt hij onder de werkloze beroepsbevolking.
Anders valt hij onder de niet-beroepsbevolking.
Zie voor een schematisch overzicht van de bepaling werkloosheid de *Barometer beroepsbevolking*.
1 = werkzame beroepsbevolking; 2 = werkloze beroepsbevolking; 3 = niet beroepsbevolking; 7 = n.v.t.

5. Arbeidsduur per week

Het antwoord op de vraag: hoeveel uur werkt u in totaal gemiddeld per week, overuren en onbetaalde uren niet meegerekend?
1 = minder dan 12 uur; 2 = 12-20 uur; 3 = 20-35 uur; 4 = 35 uur of meer.

6. Meer of minder willen werken

Hier wordt aan de respondenten die tot de beroepsbevolking behoren gevraagd of ze meer of minder willen werken dan hun arbeidsduur per week.
1 = meer willen werken; 2 = minder willen werken; 3 = niet meer/minder willen werken;
7 = vraag niet gesteld.

7. Bereidheid tot werken

Hier wordt aan de respondenten die tot de niet-beroepsbevolking of werkloze beroepsbevolking behoren gevraagd of ze wel of niet 12 uur of meer per week zouden willen en kunnen werken. Het minimum van 12 uur wordt voor de Nederlandse definitie van werkloosheid aangehouden als indicatie dat de respondent een substantieel aantal uren per week werkt. Vanaf anderhalve dag per week geeft namelijk de meerderheid van de mensen aan dat betaald werken de belangrijkste bezigheid is.
1 = persoon wil niet werken/wel werken, maar kan niet werken;
2 = persoon wil minder dan 12 uur per week werken/wil wel 12 uur of meer per week werken, maar kan niet;
3 = persoon wil 12 uur of meer per week werken/persoon heeft werk gevonden van 12 uur of meer per week;
7 = n.v.t.

8. Beroepsrichting

Hier is voor respondenten die tot de werkzame beroepsbevolking behoren afgeleid in welke categorie zijn beroep of functie het beste past.
00 = Geen werkkring; 02 = Docenten en staffuncties onderwijs, onderwijskundig;
04 = Agrarisch/Exact; 06 = Technisch; 08 = Transport, communicatie en verkeer;
08 = Medisch en paramedisch; 10 = Economisch, administratief en commercieel;
13 = Juridisch, bestuurlijk en openbare orde en veiligheid/Taal en cultuur;
16 = Gedrag en maatschappij; 17 = Persoonlijke en sociale verzorging; 18 = Management/Algemeen;
99 = Beroepsrichting onbekend.

9. Onderwijsniveau

In deze variabele is het hoogst behaalde onderwijsniveau van de respondent afgeleid:
1 = Laag: basisonderwijs, lbo, vbo, vso, vmbo, mavo, ulo, mulo;
2 = Midden: havo, mms, vwo, hbs, mbo;
3 = Hoog: hbo, wo.

10. Publicatie indeling onderwijsrichting

In deze variabele is de richting van het hoogst behaalde onderwijsniveau van de respondent afgeleid.
1 = Algemeen; 2 = Leraren; 3 = Humaniora, sociale wetenschap, communicatie en kunst;
4 = Economie, commercieel, management en administratie;
5 = Juridisch, bestuurlijk, openbare orde en veiligheid; 6 = Wiskunde, natuurwetenschap en informatica;
7 = Techniek; 8 = Agrarisch en milieu; 9 = Gezondheidszorg, sociale dienstverlening en verzorging;
10 = Horeca, toerisme, vrijetijdsbesteding, transport en logistiek; 11 = Onbekend.

LES 1
Tabellen en grafieken maken

Eerst moet het gegevensbestand worden ingelezen:

- (Ga naar digiboek4).
- Ga naar VU-Statistiek.
- Data analyse.
- Open bestand.
- Open het bestand ‘beroepsbevolking.vus’.
- Controleer of het aantal records 76746 is (zie links onderin).
- Controleer dat er 10 variabelen in het bestand zitten via ‘Data’, ‘Variabelen’.

Alle variabelen in het bestand zijn gemeten op nominaal niveau of ordinaal niveau.

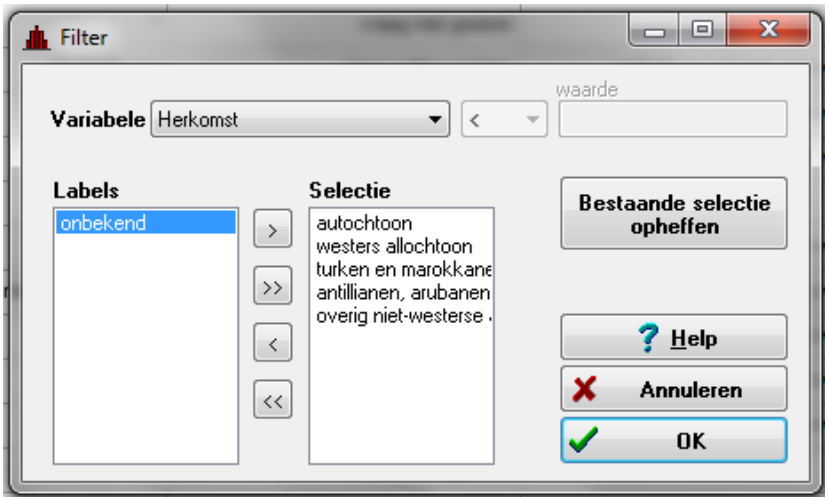
a. Geef van elke variabele in het bestand het meetniveau (nominaal of ordinaal).
b. Maak een staafgram voor de verschillende leeftijdsgroepen. Splits deze op geslacht.
c. Hoeveel procent is hoogopgeleid?
   Is er een verschil tussen mannen en vrouwen voor wat betreft hun opleidingsniveau?
d. Hoeveel procent is westers allochtoon?
e. Hoeveel procent van de mannen uit de beroepsbevolking wil meer werken?
f. Maak een staafdiagram van de verdeling van de arbeidsduur per week
   Is deze voor jongeren anders dan voor ouderen?
g. Hebben mensen die in het onderwijs werken ook vaak een opleiding in die richting gehad?

LES 2
Bestand opschonen en aanpassen

Een groot gegevensbestand is in de praktijk meestal niet precies gemaakt zoals je het wilt hebben voor je eigen onderzoek. Vaak moet je een bestand eerst ‘opschonen’ en/of de variabelen ‘aanpassen’.
In deze les gaan we in op een aantal in de praktijk veel voorkomende handelingen bij het werken met grote gegevensbestanden.

Filteren/selecteren

In een bestand zitten vaak records waarvan bij een of meerdere variabelen de waarde onbekend is. In VuStat kun je deze records (tijdelijk) verwijderen. Dit noemen we filteren.
Bijvoorbeeld, bij de variabele *herkomstgroepering* komt de waarde 'onbekend' voor. Je kunt de betreffende records tijdelijk verwijderen door het hanteren van een filter.
Ga naar 'Data', 'Selectiefilter', 'Filter' en plaats alle labels in de selectie behalve het label 'onbekend'.



Druk vervolgens op OK. Het resultaat is een selectie van 76676 records (zie links onderin).

Je kunt een aangebracht filter opheffen door gebruik te maken van de mogelijkheid 'Selectiefilter opheffen'. Doe dit.

a. We willen nagaan hoe vaak mensen die werkzaam zijn in het onderwijs aangeven dat zij minder willen werken. Gebruik een selectiefilter om de betreffende personen te selecteren en maak vervolgens een frequentietabel van de variabele *meer of minder willen werken*.

Hercoderen

In een bestand zitten soms kwalitatieve variabelen met verschillende mogelijke waarden waarvan je sommige waarden wilt samenvoegen. Je kunt dan de betreffende variabele hercoderen.
Bijvoorbeeld, de variabele herkomstgroepering heeft als mogelijke waarden: autochtonen (10), westerse allochtonen (20), Turken en Marokkanen (31), Antillianen/Arubanen en Surinamers (33), overige niet-westerse allochtonen (35) en onbekend (99). Voor sommige berekeningen zijn we wellicht alleen geïnteresseerd in het onderscheid tussen autochtonen en de rest. Dan zouden we dus eerst de personen met herkomstgroepering onbekend eruit kunnen filteren en vervolgens de personen met herkomstgroepering 20, 31, 33, en 35 kunnen samenvoegen door een hercodering.

Ga naar ‘Data’, ‘Hercoderen’, bij bronvariabele *Herkomst* invullen, bij waarde ‘10’, bij doelvariabele ‘nieuwe variabele aanmaken’, geef deze de naam *herkomst2*, geef aan dat het soort variabele labels is en geef het getal 0 de labeltekst ‘autochtoon’ en het getal 1 de labeltekst ‘allochtoon’.
Je krijgt dan het volgende scherm:

Vul bij waarde ‘0’ in en druk op ‘Voegtoe’. Er wordt dan een regel opgenomen in de lijst met codeformules.

Geef vervolgens bij bronvariabele de waarde ‘20’ en bij doelvariabele de waarde ‘1’ en ‘Voegtoe’.

Herhaal dit voor bronvariabele waarde 31, 33 en 35 en geef steeds doelvariabele de waarde ‘1’.
Sluit af met OK.

Kijk nu in de lijst met variabelen (via ‘Data’, ‘Variabelen’).
Als het goed is, staat daarin nu een nieuwe variabele, namelijk *herkomst2*.

b. Voer voorgaande hercodering uit en maak daarna een frequentietabel van beroepsbevolking. Splits deze tabel op *herkomst2*. Dit doe je via ‘Splitsen’, splitsvariabele is *herkomst2.*
Gebruik de optie procenten en lees af welk percentage van de autochtonen tot de werkloze beroepsbevolking hoort en welk percentage van de allochtonen tot de werkloze beroepsbevolking hoort.

In het bestand zit de variabele *leeftijdsgroep*. Hierbij is de leeftijd ingedeeld in klassen met een klassebreedte van 10 jaar. De variabele *leeftijdsgroep* zoals die in het bestand is opgenomen heeft dus het ordinale meetniveau.
Voor sommige berekeningen zouden we natuurlijk liever beschikken over de leeftijd in jaren, met andere woorden de leeftijd als een kwantitatieve variabele. We kunnen deze situatie nabootsen door het klassenmidden van de leeftijdsklassen te hanteren.
Ga naar ‘Data’, ‘Hercoderen’, vul bij bronvariabele *Leeftijdsgroep* in, en maak bij doelvariabele een nieuwe variabele aan. Geef de nieuwe variabele de naam *leeftijd*. Kies bij soort variabele voor een geheel getal. Als je hier kommagetal kiest, loop je het risico dat VuStat vastloopt.
Vul bij bronvariabele de waarde ‘= 1’ in en bij doelvariabele de waarde ‘20’. Klik vervolgens op ‘Voegtoe’. Je krijgt het onderstaande scherm.

Leeftijdsgroep 2 moet leeftijd 30 krijgen, leeftijdsgroep 3 leeftijd 40, leeftijdsgroep 4 leeftijd 50 en leeftijdsgroep 5 leeftijd 60.

Voor kwantitatieve variabelen is het handig een overzicht te hebben van de zogenaamde kentallen. Dit zijn de centrum- en spreidingsmaten van de variabele. In VuStat kun je deze uitrekenen via kentallen.

c. Voer bovenstaande hercodering uit voor *leeftijdsgroep* en bereken vervolgens de kentallen van leeftijd.

In het vervolg van dit computerpracticum heb je een hercodering nodig van de variabele *arbeidsduur*. Deze hercodering ga je hier alvast maken.
Hercodeer de variabele *arbeidsduur* als volgt:

- Als de arbeidsduur minder 12 uur is, maak er 6 van.
- Als de arbeidsduur van 12 tot 20 uur is, maak er 16 van.
- Als de arbeidsduur van 20 tot 35 uur is, maak er 28 van.
- Als de arbeidsduur 35 uur of meer is, maak er 38 van.

Geef de nieuwe variabele de naam *arbeidsduur2* en gebruik als soort variabele een geheel getal. Als je kommagetal gebruik dan loop je het risico dat VuStat foutmeldingen gaat geven.

d. Voer bovenstaande hercodering uit voor *arbeidsduur* en bereken vervolgens de kentallen van arbeidsduur.

In het vervolg van deze lessenserie heb je het VuStatbestand nodig waarin bovenstaande hercoderingen al zijn uitgevoerd (beroepsbevolking_2.VUS).

LES 3
Betrouwbaarheidsintervallen berekenen

Op basis van bestanden zoals de enquête beroepsbevolking kunnen we uitspraken doen over populatieproporties en populatiegemiddelden, bijvoorbeeld over de proportie werklozen binnen de beroepsbevolking in Nederland of over de gemiddelde leeftijd van de personen die werkzaam zijn in het onderwijs in Nederland.
Omdat het gegevensbestand behoorlijk groot is (veel personen bevat), kunnen we met grote mate van zekerheid vrij nauwkeurige uitspraken doen. En dat is natuurlijk precies wat beleidsmakers willen, want aan onzekere en/of onnauwkeurige uitspraken hebben zij niet veel.

a. Bereken het 95%-betrouwbaarheidsinterval voor de proportie werklozen binnen de beroepsbevolking.
b. Bereken het 95%-betrouwbaarheidsinterval voor de gemiddelde leeftijd van de personen die werkzaam zijn in het onderwijs. Gebruik hiervoor de variabele *leeftijd*, een hercodering van de variabele *leeftijdsgroep*.

LES 4
Verschil tussen twee groepen berekenen

Er zijn verschillende manieren om het verschil tussen twee groepen te duiden:
is er een gering, middelmatig of groot verschil?
In de media kun je regelmatig lezen dat met name laagopgeleiden getroffen worden door de economische crisis, omdat zij een groter risico lopen om werkloos te geraken dan hoogopgeleiden.
In dit kader ligt het voor de hand om na te gaan hoe groot het verschil is in proportie werklozen tussen laag- en hoogopgeleiden.

a. Maak een 2x2-kruistabel en bereken phi. Wat kun je concluderen op basis van phi?

In de media staat ook regelmatig dat het onderwijs vergrijst. Daarmee wordt bedoeld dat de mensen die werkzaam zijn in het onderwijs gemiddeld genomen steeds ouder worden. Dit kan een probleem gaan vormen op het moment er geen (jonge) nieuwe docenten zijn om docenten die de pensioengerechtigde leeftijd bereiken te vervangen.
In dit licht kun je je afvragen hoe groot het verschil is in leeftijdsopbouw tussen personen die werkzaam zijn in het onderwijs en personen die een andere beroepsrichting hebben.

b. Bereken het maximale cumulatieve percentageverschil om na te gaan hoe groot het verschil in leeftijd is tussen personen die werkzaam zijn in het onderwijs en de personen die een andere beroepsrichting hebben.
   Wat kun je concluderen op basis van het maximale cumulatieve percentageverschil?

Het onderwijs schijnt tevens te feminiseren. Oftewel, er zijn steeds meer vrouwen in het onderwijs werkzaam zijn ten opzichte van mannen. Wel werken vrouwen vaker parttime en mannen vaker fulltime, dus misschien valt het allemaal wel mee met die feminisering.
In dit verband kun je je afvragen hoe groot het verschil is in arbeidsduur tussen mannen en vrouwen die werkzaam zijn in het onderwijs. Daartoe ga je de effectgrootte uitrekenen. Daarbij maak je gebruik van de variabele *arbeidsduur2*, een hercodering van de variabele *arbeidsduur*.

c. Bereken vervolgens de effectgrootte om na te gaan hoe groot het verschil is tussen de arbeidsduur van mannen die werkzaam zijn in het onderwijs en vrouwen die werkzaam zijn in het onderwijs.

LES 5
Samenhang tussen twee kwantitatieve variabelen onderzoeken

In deze les onderzoeken we de mate van samenhang tussen de variabelen *leeftijd* en *arbeidsduur2*. Je gebruikt dus de hercoderingen van de variabelen *leeftijdsgroep* en *arbeidsduur*.

Maak vervolgens de puntenwolk met ‘leeftijd in jaren’ op de x-as en ‘arbeidsduur in uren’ op de y-as.
Wat is de formule van de trendlijn? Welke voorspelling van de arbeidsduur in uren geeft de trendlijn voor een persoon met een leeftijd van 30 jaar? En voor iemand van 58 jaar?

LES 6
Onderzoek 1: werkloosheid

Het ging enige jaren minder goed met de economie in ons land. In de media verschenen veel berichten over de economische crisis en de gevolgen ervan. Men maakte zich zorgen over de grote werkloosheid onder laagopgeleide jongeren.
In deze les ga je onderzoeken hoe werkloosheid samenhangt met leeftijd en opleidingsniveau.
We kijken enkel naar de beroepsbevolking, dus selecteer in de beroepsbevolking: werkzame en werkloze beroepsbevolking.

a. Bereken hoeveel procent van de beroepsbevolking in 2011 werkloos is.
b. Ga na of in de beroepsbevolking jongeren (15-24 jaar) vaker werkloos zijn dan ouderen (55-64 jaar).
c. Ga na of er in de beroepsbevolking onder laagopgeleiden een grotere werkloosheid heerst dan onder hoogopgeleiden.
d. Onderzoek met behulp van phi hoe groot het verschil is in werkloosheid tussen laagopgeleide jongeren en hoogopgeleide ouderen.

LES 7
Onderzoek 2: opleidingsniveau

Iemand doet de volgende uitspraken over opleidingsniveau:
- Jongeren zijn hoger opgeleid dan ouderen.
- Mannen zijn hoger opgeleid dan vrouwen, maar dit verschil wordt steeds kleiner.

Ga na of de informatie in het gegevensbestand aanleiding geeft om deze uitspraken te ondersteunen of juist te verwerpen.
Gebruik geschikte tabellen, grafieken, kentallen, betrouwbaarheidsintervallen en/of verschilmaten.

§ 4.10 Diagnostische computertoets

Deze toets duurt 90 minuten. De toets bestaat uit negen onderdelen. Voor elk onderdeel staat hoeveel punten je er maximaal mee kunt behalen. In totaal kun je voor deze toets maximaal 70 punten behalen.

In deze toets maak je gebruik van het gegevensbestand enquête beroepsbevolking 2011 (beroepsbevolking_2.vus).
In dit bestand staan de variabelen *leeftijd* en *arbeidsduur2*, de hercoderingen van *leeftijdsgroep* en *arbeidsduur*. Je mag in de toets de variabelen *leeftijd* en *arbeidsduur2* gebruiken als kwantitatieve variabelen. De oorspronkelijke variabelen *leeftijdsgroep* en *arbeidsduur* zijn kwalitatief en van het ordinale meetniveau.

Het bestand telt 76746 records.
Het bestand bevat onderstaande variabelen.
Je kunt de gebruikte codes in VuStat zichtbaar maken door te klikken op de knop 'Labels'.

1. Geslacht
1 = man; 2 = vrouw

2. Leeftijdsgroep
Leeftijd wordt berekend naar aanleiding van geboortedatum en vervolgens ingedeeld in tienjaarsklassen: 1 = 15-24 jaar; 2 = 25-34 jaar; 3 = 35-44 jaar; 4 = 45-54 jaar; 5 = 55-64 jaar.

3. Herkomstgroepering
De herkomstgroepering wordt bepaald naar aanleiding van het geboorteland van beide ouders.
Als beide ouders in Nederland zijn geboren is de respondent autochtoon. Als een van beide ouders in het buitenland is geboren is de respondent allochtoon. Wanneer beide ouders in een ander land zijn geboren, kijkt het CBS naar het geboorteland van de moeder om de herkomstgroepering te bepalen. Voorbeeld: moeder is geboren in Turkije, vader in Duitsland, dan heeft de respondent Turkije als herkomstland.
10 = autochtonen; 20 = westerse allochtonen; 31 = Turken en Marokkanen;
33 = Antillianen/Arubanen en Surinamers'; 35 = overig niet-westerse landen; 99 = onbekend.

4. Beroepsbevolking
Heeft de respondent voor 12 uur per week of meer werk? Dan valt hij onder de werkzame beroepsbevolking.
Wil of kan de respondent niet werken voor 12 uur of meer per week (bijvoorbeeld scholieren en huisvrouwen/-mannen)? Dan valt hij onder de niet-beroepsbevolking.
Wil de respondent voor 12 uur of meer werken en doet hij dat nu niet? Én kan hij op korte termijn beginnen? Én zoekt hij actief naar werk? Dan valt hij onder de werkloze beroepsbevolking. Anders valt hij onder de niet-beroepsbevolking.
Zie voor een schematisch overzicht van de bepaling werkloosheid de *Barometer beroepsbevolking*.
1 = werkzame beroepsbevolking; 2 = werkloze beroepsbevolking; 3 = niet beroepsbevolking; 7 = n.v.t.

5. Arbeidsduur per week
Het antwoord op de vraag: hoeveel uur werkt u in totaal gemiddeld per week, overuren en onbetaalde uren niet meegerekend?
1 = minder dan 12 uur; 2 = 12-20 uur; 3 = 20-35 uur; 4 = 35 uur of meer.

6. Meer of minder willen werken

Hier wordt aan de respondenten die tot de beroepsbevolking behoren gevraagd of ze meer of minder willen werken dan hun arbeidsduur per week.
1 = meer willen werken; 2 = minder willen werken; 3 = niet meer/minder willen werken;
7 = vraag niet gesteld.

7. Bereidheid tot werken

Hier wordt aan de respondenten die tot de niet-beroepsbevolking of werkloze beroepsbevolking behoren, gevraagd of ze wel of niet 12 uur of meer per week zouden willen en kunnen werken.
Het minimum van 12 uur wordt voor de Nederlandse definitie van werkloosheid aangehouden als indicatie dat de respondent een substantieel aantal uren per week werkt. Vanaf anderhalve dag per week geeft namelijk de meerderheid van de mensen aan dat betaald werken de belangrijkste bezigheid is.
1 = persoon wil niet werken / wel werken, maar kan niet werken;
2 = persoon wil minder dan 12 uur per week werken / wil wel 12 uur of meer per week werken maar kan niet;
3 = persoon wil 12 uur of meer per week werken / persoon heeft werk gevonden van 12 uur of meer per week;
7 = n.v.t.

8. Beroepsrichting

Hier is voor respondenten die tot de werkzame beroepsbevolking behoren afgeleid in welke categorie zijn beroep of functie het beste past.
00 = Geen werkkring; 02 = Docenten en staffuncties onderwijs, onderwijskundig;
04 = Agrarisch/Exact; 06 = Technisch; 08 = Transport, communicatie en verkeer;
09 = Medisch en paramedisch; 10 = Economisch, administratief en commercieel;
13 = Juridisch, bestuurlijk en openbare orde en veiligheid/Taal en cultuur;
16 = Gedrag en maatschappij; 17 = Persoonlijke en sociale verzorging; 18 = Management/Algemeen;
99 = Beroepsrichting onbekend.

9. Onderwijsniveau

In deze variabele is het hoogst behaalde onderwijsniveau van de respondent afgeleid:
1 = Laag: basisonderwijs, lbo, vbo, vso, vmbo, mavo, ulo, mulo.
2 = Midden: havo, mms, vwo, hbs, mbo.
3 = Hoog: hbo, wo.

10. Publicatie indeling onderwijsrichting

In deze variabele is de richting van het hoogst behaalde onderwijsniveau van de respondent afgeleid.
1 = Algemeen; 2 = Leraren; 3 = Humaniora, sociale wetenschap, communicatie en kunst; 4 = Economie, commercieel, management en administratie; 5 = Juridisch, bestuurlijk, openbare orde en veiligheid;
6 = Wiskunde, natuurwetenschap en informatica; 7 = Techniek; 8 = Agrarisch en milieu;
9 = Gezondheidszorg, sociale dienstverlening en verzorging; 10 = Horeca, toerisme, vrijetijdsbesteding, transport en logistiek; 11 = Onbekend.

Eerst moet het gegevensbestand worden ingelezen:

- (Ga naar digiboek4).
- Ga naar VU-Statistiek.
- Data analyse.
- Open bestand.
- Open het bestand ‘beroepsbevolking.vus’.
- Controleer of het aantal records 76746 is (zie links onderin).
- Controleer dat er 10 variabelen in het bestand zitten via ‘Data’, ‘Variabelen’.

In deze toets kijken we naar mensen die werkzaam zijn in de sector gezondheidszorg. Je kunt deze personen selecteren door een selectiefilter op beroepsrichting gelijk aan 8. Doe dit.
De eerste drie vragen kun je beantwoorden met behulp van enkele eenvoudige tabellen, grafieken en/of kentallen.

1. Hoeveel procent van de mensen is autochtoon?
2. Wat is de gemiddelde arbeidsduur en wat is de standaardafwijking?
3. Hoeveel procent van de mensen is vrouw en jonger dan 35 jaar?

We nemen aan dat het bestand een representatieve steekproef is van de totale bevolking tussen 15 en 65 jaar. Op basis van deze steekproef kun je dan uitspraken doen over populatieproporties en populatiegemiddelden

4. Bereken het 95%-betrouwbaarheidsinterval voor de proportie vrouwen werkzaam in de gezondheidszorg.
5. Bereken het 95%-betrouwbaarheidsinterval voor de gemiddelde leeftijd van de mensen die werkzaam zijn in de gezondheidszorg.

In het vervolg van deze toets kijk je naar het verschil tussen mannen en vrouwen die werkzaam zijn in de gezondheidszorg. De bedoeling is dat je drie uitspraken controleert:
Er is een groot verschil tussen mannen en vrouwen voor wat betreft:

a. arbeidsduur;
b. onderwijsniveau;
c. proportie die meer zou willen werken.

6. Controleer deze drie uitspraken.

Het gegevensbestand bevat variabelen die een samenhang zouden kunnen vertonen. In het vervolg van deze toets zie je zo’n samenhang terug in het bestand.
Een logische gedachte lijkt dat juist mensen met een fulltime baan minder zouden willen werken en dat juist mensen met een parttime baan meer zouden willen werken.

7. Onderzoek met behulp van een kruistabel of de variabelen *arbeidsduur* en *meer of minder willen werken* op deze manier samenhangen.

Het is op voorhand niet duidelijk of arbeidsduur positief samenhangt met leeftijd of juist negatief. Voor beide is op voorhand wel wat te zeggen dat steeds meer mensen parttime werken in plaats van fulltime, dus jongeren werken wellicht vaker parttime en ouderen vaker fulltime. Aan de andere kant is het denkbaar dat vanwege de zwaarte van het werk de ouderen niet langer fulltime werken maar parttime en dat de jongeren minder moeite hebben met de zwaarte van het werk en dus vaker fulltime werken.

8. Onderzoek met behulp van een puntenwolk de samenhang tussen *leeftijd* en *arbeidsduur*.

Voor het laatste onderdeel van deze toets heb je het gehele bestand van 76746 records nodig.
Je moet dus eerst het bestaande selectiefilter opheffen. Doe dit.

Vacatures in de gezondheidszorg blijken soms moeilijk vervuld te kunnen worden. Sommige mensen beweren dat er wel een groot arbeidspotentieel is omdat met name veel vrouwen enerzijds wel een opleiding hebben genoten in de richting van de gezondheidszorg. maar anderzijds niet (meer) tot de beroepsbevolking gerekend worden.

9. Onderzoek hoe groot deze groep is.