Statistiek met de computer

8

Voorkennis Werken met Excel

Theorie A Data invoeren

Een **spreadsheet** of **(digitaal) rekenblad** bestaat uit rijen, kolommen en cellen. Elke cel kan een getal, tekst of een formule bevatten. Formules maken het mogelijk om bijvoorbeeld bij een kolom het gemiddelde te bereken en dit weer te geven in een cel. Een spreadsheet is dynamisch. Dat wil zeggen dat de uitkomsten van de formules direct opnieuw worden berekend als je de waarde in een cel verandert.
In dit hoofdstuk werken we met het spreadsheetprogramma Excel. De Excelmodules en de bestanden die in dit hoofdstuk genoemd worden vind je in *Getal & Ruimte online*.

[►EXCEL] In de Excelmodule **Data invoeren** wordt behandeld

- wat rijen, kolommen en cellen zijn en wat bedoeld wordt met de termen rekenboek, rekenblad en tabblad
- hoe je handig data invoert in rijen en kolommen
- hoe je cellen selecteert, de inhoud van cellen verplaatst en de vulgreep gebruikt
- hoe je cellen opmaakt.

Neem deze module door.

Bijzondere muiscursors in Excel


selectiecursor


verplaatscursor


vulgreep

1 De tabel hieronder gaat over de aantallen verkochte woningen in de gemeente Utrecht. Open een nieuw rekenblad en voer de tabel in. Gebruik zowel in rij 1 als in kolom A de vulgreep. Zorg ook voor de opmaak zoals in de tabel hieronder. Sla het bestand op onder de naam WoningVerkoopUtrecht.xlsx. Je hebt het bestand later weer nodig.

	A	B	C	D	E
1	jaar	1e kwartaal	2e kwartaal	3e kwartaal	4e kwartaal
2	2003	877	959	1098	1182
3	2005	894	1061	1319	1319
4	2007	1035	1123	1292	1252
5	2009	741	743	881	964
6	2011	735	767	864	889
7	2013	544	484	710	882

figuur 8.1

Theorie B Rekenen met Excel

[►EXCEL] In de Excelmodule **Formules en verwijzingen** leer je wat absolute verwijzingen en wat relatieve verwijzingen naar cellen zijn en hoe je deze gebruikt in formules.

Neem deze module door.

Gebruik F4 voor het wisselen tussen absolute en relatieve verwijzingen
B2 = relatieve verwijzing naar cel B2
B2 = absolute verwijzing naar cel B2
B$2 = relatieve kolom B, de rij blijft 2
$B2 = de kolom blijft B en rij 2 is relatief

2 In het bestand ProefwerkScores.xlsx zie je voor elke leerling van een klas hoeveel punten hij of zij gehaald heeft voor elk van de vragen van een proefwerk. In rij 2 zie je per vraag hoeveel punten er maximaal gescoord konden worden.

a Bereken in cel Y2 hoeveel punten er maximaal gescoord konden worden bij dit proefwerk. Welke Excelfunctie heb je gebruikt?

b Bereken in cel Y3 hoeveel punten Joris gescoord heeft bij dit proefwerk. Gebruik de vulgreep in kolom Y om de scores van de andere leerlingen te berekenen. Welke leerling heeft de meeste punten gehaald en welke leerling de minste?

Belangrijke Excelfuncties
SOM
AANTAL
GEMIDDELDE
MEDIAAN
MIN
MAX
AFRONDEN
STDEV.P
KWARTIEL

Het cijfer voor dit proefwerk wordt berekend met de formule

$$\text{cijfer} = \frac{\text{score}}{\text{max score}} \times 9 + 1.$$

c Bereken in cel Z3 het cijfer dat Joris voor dit proefwerk heeft gehaald. Gebruik in je formule een absolute verwijzing naar cel Y2 en een relatieve verwijzing naar cel Y3.

d Pas de formule in Z3 aan zodat het cijfer wordt afgerond op één decimaal. Gebruik hiervoor de functie AFRONDEN. Gebruik de vulgreep om de cijfers van de andere leerlingen te berekenen.

De docent is geïnteresseerd in zowel statistieken per vraag als statistieken van de totaalscores en cijfers van de leerlingen.

e Bereken voor elke vraag het gemiddelde, de standaardafwijking, de laagste score, Q1, de mediaan, Q3 en de hoogste score. Doe dit ook voor de totaalscores en de cijfers.

f Gebruik formules om de volgende vragen te beantwoorden.
- Bij welke vragen is de kwartielafstand van de scores 2,5?
- Bij welke vragen is het gemiddelde gelijk aan de mediaan?

3 In het bestand Cijfers4Ha.xlsx zie je een overzicht van de proefwerkcijfers van klas 4Ha. Het eindcijfer wordt bepaald door een gewogen gemiddelde te nemen van de cijfers. In rij 1 zie je hoe vaak elk proefwerk meetelt voor het eindcijfer. Het eindcijfer wordt afgerond op één decimaal.

a Bereken voor elke leerling het eindcijfer.

b Bereken in cel J25 het gemiddelde eindcijfer van de klas. Pas de celeigenschappen van deze cel aan zodat het gemiddelde wordt weergegeven in twee decimalen.

c Als de laatste toets niet 5 maar 6 keer meetelt dan is het gemiddelde eindcijfer van de klas hoger. Hoeveel scheelt het?

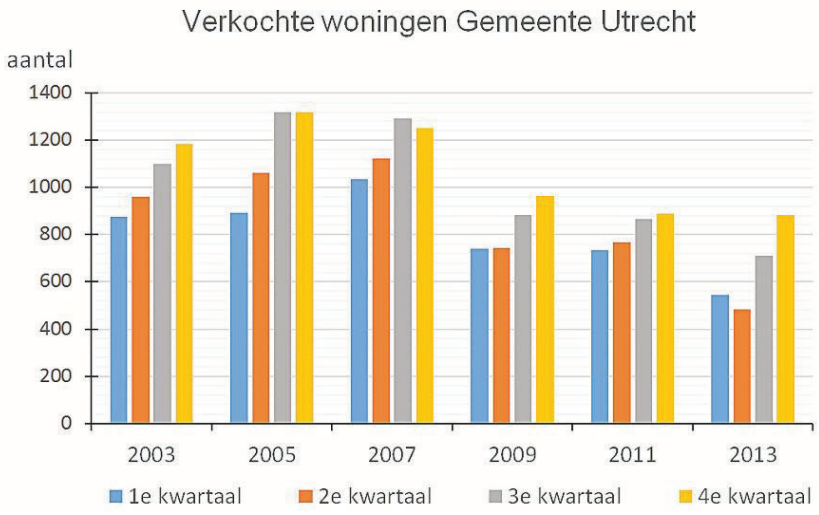
Afronden
- Met celeigenschappen: alle decimalen blijven behouden.
- Met de functie AFRONDEN: de overige decimalen gaan verloren.

Theorie C Diagrammen

Onderzoeksgegevens worden vaak verzameld in tabellen. Tabellen geven echter niet altijd een goed overzicht van de gegevens. Daarom worden bij de tabellen diagrammen gemaakt. Een diagram vat de gegevens op een grafische manier samen. Bekende diagrammen zijn lijn-, staaf-, stapel- en cirkeldiagrammen. Een diagram moet snel inzicht geven in de gegevens en moet dus gemakkelijk te lezen zijn. Zorg bij het maken van diagrammen dus voor
- een passende grafiektitel
- duidelijke informatie bij de assen en/of een legenda
- een geschikte schaalverdeling met eventueel roosterlijnen
- logisch kleurgebruik.

[►EXCEL] In de Excelmodule **Diagrammen in Excel** wordt uitgelegd hoe je lijn-, staaf-, stapel- en cirkeldiagrammen maakt.

4 Open het bestand WoningVerkoopUtrecht.xlsx van opgave 1 en maak bij de gegevens het samengestelde staafdiagram hiernaast.



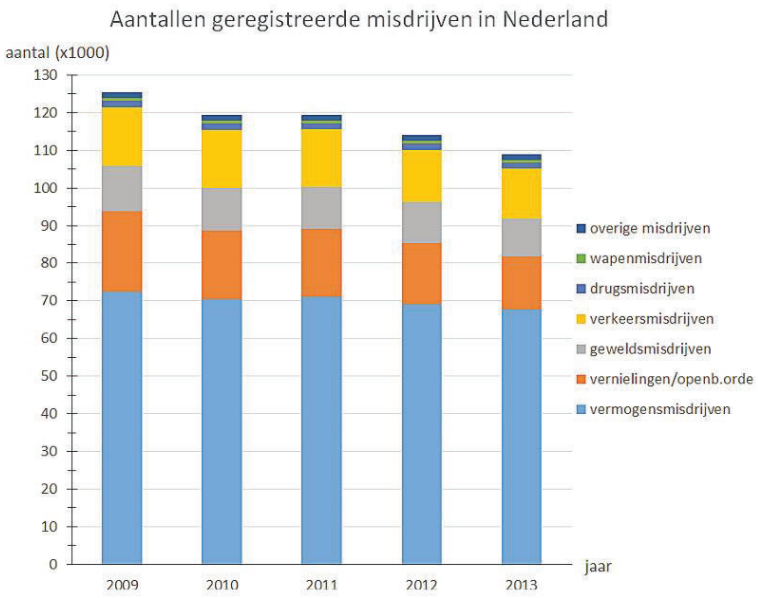
figuur 8.2

5 In het bestand Criminaliteit.xlsx staan cijfers over geregistreerde criminaliteit in Nederland in de jaren 2003 tot en met 2013.

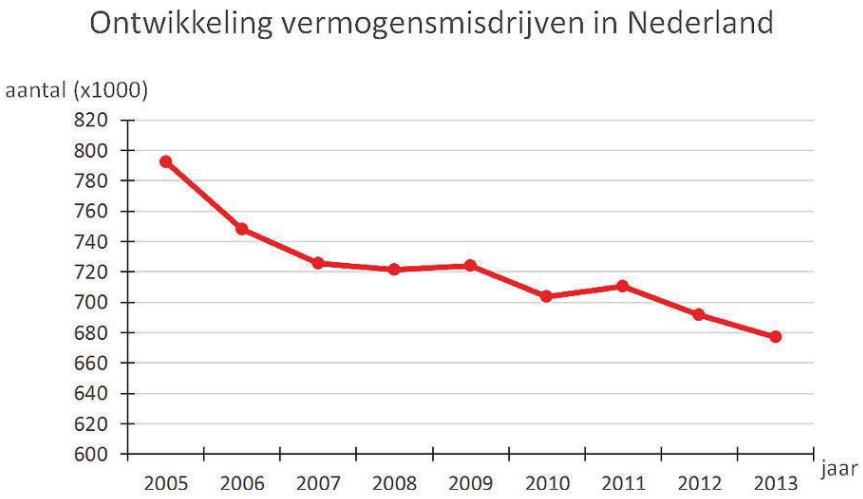
a Maak voor de jaren 2009 tot en met 2013 een stapeldiagram zoals in figuur 8.3.

b Het lijndiagram in figuur 8.4 geeft de ontwikkeling weer van het aantal vermogensmisdrijven in de periode 2005-2013.
Maak dit diagram.

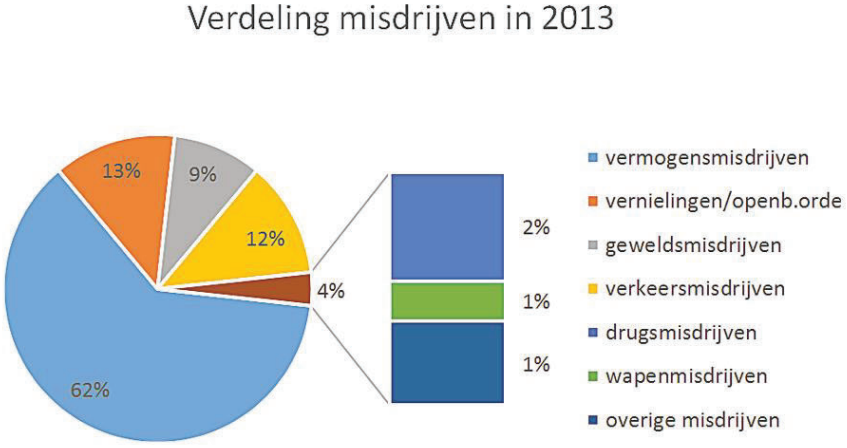
c Het diagram in figuur 8.5 geeft weer hoe de in 2013 gepleegde misdrijven zijn verdeeld over de verschillende categorieën.
Maak dit diagram.



figuur 8.3



figuur 8.4



figuur 8.5

8

8.1 Werken met datasets

Theorie A Exceltabellen, filteren en sorteren

Om zonder problemen gebruik te kunnen maken van een aantal geavanceerde mogelijkheden van Excel is het belangrijk dat je werkt met een zogenaamde **nette dataset**. Bij een **nette dataset** zijn de kolommen voorzien van kolomkoppen en zijn de cellen die grenzen aan de dataset leeg. Bovendien staan gegevens die bij elkaar horen steeds in één rij. De informatie in zo'n rij heet een **record**. Zo zie je in figuur 8.6 vier records.
Een nette dataset kun je omzetten in een zogenaamde **Exceltabel**. Een Exceltabel geeft je onder andere de mogelijkheid gegevens te **sorteren** en te **filteren**. Bovendien kun je bij een Exceltabel in een **totalenrij** snel bijvoorbeeld het gemiddelde berekenen van een deelgroep.

[►EXCEL] In de module **Exceltabellen** leer je hoe je een nette dataset omzet in een Exceltabel, hoe je in Exceltabellen sorteert, filtert en werkt met de totalenrij.

Neem de module door.

In figuur 8.6 zie je een Exceltabel van de Formule 1 resultaten over de eerste helft van 2014. Je ziet de variabelen coureur, nationaliteit, team en punten. De kolom 'team' is gefilterd op 'Ferrari' en 'McLaren-Mercedes'. De kolom 'punten' is gesorteerd van hoog naar laag en in de totalenrij is de functie SOM gebruikt.

	A	B	C	D
1	coureur	nationaliteit	team	punten
4	Fernando Alonso	Spaans	Ferrari	61
9	Jenson Button	Brits	McLaren-Mercedes	31
10	Kevin Magnussen	Deens	McLaren-Mercedes	21
13	Kimi Räikkönen	Fins	Ferrari	17
24	**Totaal**			**130**

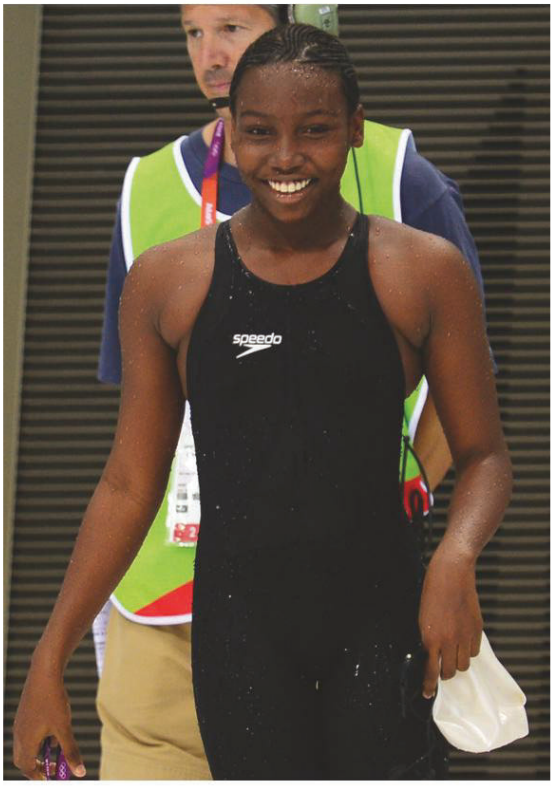
figuur 8.6 Formule 1 resultaten tot 1 juni 2014

Afspraak
Geef bij de beantwoording van de vragen steeds aan
- in welke kolommen je welke filters gebruikt
- in welke kolommen je op welke manier sorteert
- welke functies je gebruikt.

1 De volgende vragen gaan over het weer in De Bilt. Gebruik het bestand WeerHistorie.xlsx dat een nette dataset bevat met weergegevens gemeten door het weerstation in De Bilt.

a Hoeveel variabelen bevat de dataset?
b Van hoeveel dagen bevat de dataset gegevens?
c Zet de gegevens om in een Exceltabel en maak de totalenrij aan.
d Wat was de warmste dag?
e Op hoeveel dagen was het gemiddeld 0 °C of kouder?
f Een dag wordt 'tropisch' genoemd als de maximumtemperatuur 30 °C of hoger is.
Geef in één decimaal nauwkeurig de gemiddelde maximumtemperatuur van de tropische dagen.
g Het jaar 1965 was een van de natste jaren sinds de neerslagmeting.
Hoeveel mm neerslag viel er in dat jaar in totaal?
Op welke dag regende het dat jaar het langst?

2 Het bestand AtletenLonden2012.xlsx bevat een nette dataset met gegevens van 10 384 atleten die meededen aan de Olympische Spelen in 2012 in Londen.



figuur 8.7 Adzo Rebecca Kpossi uit Togo, de jongste atleet op de Olympische Spelen in 2012

a Zet de gegevens om in een Exceltabel en maak de totalenrij aan.
b Hoe oud was de jongste atleet? En hoe oud de oudste?
c Hoeveel procent van de atleten was 16 jaar of jonger?
d Wat was de gemiddelde leeftijd van de Nederlandse atleten? Rond af op gehelen.
e Wat weet je van het gewicht van de enige Nederlandse Taekwondo-deelnemer?
f De atleten die meedoen aan de triatlon hebben gemiddeld een lager gewicht dan de roeiers.
Hoeveel kg scheelt het? Rond af op gehelen.
g Volgens een journalist zijn de vrouwelijke Nederlandse atleten gemiddeld 4 cm langer dan de vrouwelijke atleten uit andere landen.
Klopt dit? En hoe zit dat bij de mannelijke atleten?

3 In het bestand StudentenEnquete.xlsx zijn gegevens verzameld van 271 studenten.

a Zet de gegevens om in een Exceltabel en maak de totalenrij aan.

b De gemiddelde lengte van de studenten in groep D is groter dan die van de studenten in groep A.
Hoeveel cm scheelt het? Rond af op gehelen.

c In welke groep is de spreidingsbreedte van de lengte het grootst en in welke groep het kleinst?

Het gemiddeld aantal uren slaap per etmaal van de studenten zie je in de kolom UrenSlaap. Studenten die gemiddeld minder dan 8 uur per etmaal slapen noemen we kortslapers.

d Hoeveel procent van de studenten is kortslaper?

e Onderzoek in welke groep het percentage kortslapers het grootst is.

f Volgens Niels zijn er relatief minder kortslapers onder de linkshandige studenten dan onder de rechtshandige studenten.
Onderzoek of Niels gelijk heeft. Licht toe.

O4 Bij een enquête onder de gasten van hotel De Negensprong is onder andere gevraagd naar de mate van tevredenheid over het ontbijtbuffet. Men kon daarbij kiezen uit de volgende antwoorden.

slecht matig voldoende ruim voldoende goed

In het bestand Ontbijtbuffet.xlsx zijn de antwoorden op deze vraag in een Exceltabel verzameld.
Bij nader inzien wil de hoteleigenaar bij zijn onderzoek alleen maar onderscheid maken tussen positieve, neutrale en negatieve beoordelingen.
Daarom wil hij de gegeven antwoorden ‘ruim voldoende’ en ‘goed’ vervangen door ‘Positief’ en de antwoorden ‘matig’ en ‘slecht’ door ‘Negatief’. De beoordeling ‘voldoende’ beschouwt hij als een neutraal antwoord en zal hij dus ‘Neutraal’ willen noemen. De hoteleigenaar bedenkt de volgende aanpak:

1 Maak een nieuwe kolom aan door in Cel D1 *Beoordeling* te typen.
2 Filter de kolom met gegeven antwoorden op ‘ruim voldoende’ en ‘goed’, zet in de eerste cel van de nieuwe kolom ‘Positief’ en gebruik de vulgreep om de kolom te vullen.
3 Filter de kolom met gegeven antwoorden op ‘matig’ en ‘slecht’, zet in de eerste cel van de nieuwe kolom ‘Negatief’ en gebruik de vulgreep om de kolom te vullen.
4 Vul op dezelfde wijze de juiste cellen van de nieuwe kolom met ‘Neutraal’.

Onderzoek of deze aanpak werkt.

Theorie B Hercoderen en formules

In opgave 4 heb je een nieuwe variabele toegevoegd die een aantal van de mogelijke waarden van een bestaande (nominale) variabele samenneemt. Dit proces wordt het **hercoderen** van een variabele genoemd.
Hercoderen van variabelen wordt vaak gedaan om onderzoeksgegevens in te delen in nieuwe categorieën. Meestal wordt bij het hercoderen een nieuwe variabele toegevoegd zodat de oorspronkelijke gegevens beschikbaar blijven. Zijn de oorspronkelijke gegevens niet meer nodig dan kun je de hercodering uitvoeren in de kolom van de oorspronkelijke gegevens.

Je kunt ook een nieuwe variabele toevoegen die waarden van verschillende bestaande variabelen combineert.
In opgave 3 heb je gebruik gemaakt van een getalfilter om onderzoek te doen onder de kortslapers. Wil je uitgebreid onderzoek doen onder de uitwonende, vrouwelijke kortslapers dan moet je nóg twee filters juist instellen. Het is dan handig te beschikken over een nieuwe variabele die aangeeft of een student al dan niet een uitwonende, vrouwelijke kortslaper is. Je hebt daarna slechts één filter nodig om deze groep te bekijken. In de volgende Excelmodule wordt uitgelegd hoe je dit doet.

[►EXCEL] Neem de Excelmodule **Hercoderen** door.

Door het combineren van variabelen in een dataset kun je vaak extra informatie verkrijgen. Zo is aan de dataset AtletenLonden2012 uit opgave 2, waarin de variabelen *Lengte* en *Gewicht* voorkomen, een nieuwe variabele *BMI* toe te voegen, die de variabelen *Lengte* en *Gewicht* combineert volgens de formule $BMI = \frac{Gewicht}{(0{,}01 \cdot Lengte)^2}$.
En met de variabelen *Goud, Zilver* en *Brons* is een nieuwe variabele *Medailles* te maken die informatie geeft over het totaal aantal medailles dat een atleet heeft gewonnen.
Behalve voor het verkrijgen van extra informatie kunnen formules ook worden gebruikt om te controleren of de gegevens in een dataset betrouwbaar zijn.

5 Ga uit van de gegevens in de dataset AtletenLonden2012.xlsx.
Voeg een variabele *HockeyDameBenelux* toe die met de waarden Ja/Nee aangeeft of een atleet hockeyster uit de Benelux (België, Nederland en Luxemburg) is.
Denk aan de afspraak op bladzijde 128.

Geef het in je uitwerking aan als je een nieuwe variabele toevoegt.

6 Ga uit van de gegevens in de dataset AtletenLonden2012.xlsx.

a Voeg een variabele *Medailles* toe die aangeeft hoeveel medailles een atleet in totaal heeft gewonnen.

b Hoeveel atleten wonnen meer dan één medaille? Hoeveel van hen hadden ten minste één gouden medaille gewonnen?

c Welke atleet heeft de meeste medailles gewonnen?

7 Ga uit van de gegevens in de dataset AtletenLonden2012.xlsx. In deze opgave bekijken we alleen atleten waarvan zowel het gewicht als de lengte bekend is. Zorg er dus voor dat je de andere atleten eruit filtert.

a Van hoeveel atleten is zowel het gewicht als de lengte bekend?

De BMI is een maat die vaak gebruikt wordt om te beoordelen of een persoon overgewicht heeft of niet. De BMI wordt berekend met de formule $BMI = \frac{gewicht}{lengte^2}$. Hierin is het gewicht in kg en de lengte in m.
Een BMI van 25 of meer geeft overgewicht aan.

b Voeg de variabele *BMI* toe. Welke atleet had de hoogste BMI?

c Voeg de variabele *OvergewichtBMI* toe die de waarde Ja heeft als de atleet overgewicht heeft en Nee als de atleet geen overgewicht heeft.

d Hoeveel atleten hadden overgewicht?

Een nadeel van de gebruikelijke formule voor de BMI is dat veel kleine mensen een te lage en veel lange mensen een te hoge BMI krijgen. Om dit op te lossen is een nieuwe formule ontwikkeld. Voor de nieuwe BMI geldt de formule $NBMI = \frac{1{,}3 \cdot gewicht}{lengte^{2{,}5}}$. Hierin is het gewicht in kg en de lengte in m.
Een NBMI van 25 of meer geeft overgewicht aan.

e Voeg de variabele *NBMI* toe aan de dataset. Welke atleet had de laagste NBMI?

f Er zijn atleten die volgens de BMI overgewicht hadden maar volgens de NBMI niet. Evenzo zijn er atleten die volgens de BMI geen overgewicht hadden maar volgens de NBMI wel. Voor hoeveel atleten geldt dit?

g Geef de lengte en het gewicht van de atleet bij wie het verschil tussen de BMI en de NBMI het grootst was.

A 8 Bij koud weer voelt het op een dag met veel wind kouder aan dan op een dag met weinig wind. Om de gevoelstemperatuur te berekenen wordt de formule $G = 13{,}12 + 0{,}6215T - 11{,}37(3{,}6W)^{0{,}16} + 0{,}3965T(3{,}6W)^{0{,}16}$ gebruikt. Hierin is T de temperatuur in °C en W de gemiddelde windsnelheid in m/s. Deze formule geldt voor $W > 1{,}3$. Als de windsnelheid 1,3 m/s of lager is, dan gaan we ervan uit dat de gevoelstemperatuur gelijk is aan de luchttemperatuur.
Ga uit van de dataset WeerHistorie.xlsx.
Onderzoek welke dag gemiddeld het koudst aanvoelde.

Als je verschillende formules in één kolom van een Exceltabel wilt gebruiken dan tik je de celverwijzingen in. Dus selecteer de cel niet met de muis.

8.2 Draaitabellen en draaigrafieken

9 In het bestand JongerenSocialeNetwerken.xlsx vind je gegevens van een enquête die in juli 2012 is gehouden onder Amerikaanse jongeren van 12 tot 18 jaar.

a Hoeveel jongens deden mee aan het onderzoek? En hoeveel meisjes?

b Geef van elk van de regio's (Midwest, Northeast, South en West) aan hoeveel jongeren meededen aan het onderzoek.

c Neem de kruistabel hiernaast over en vul deze verder in.

		geslacht M	geslacht V	
regio	Midwest			
	Northeast			
	South			
	West			
				798

Theorie A Draaitabellen maken

In opgave 9 heb je door herhaald gebruik van filters een kruistabel gemaakt. Voor deze kruistabel heb je minstens 10 keer een filter moeten instellen. Met een beetje oefening is dit snel gedaan, maar wil je een kruistabel maken waarin van elk van de staten is af te lezen hoeveel jongens en hoeveel meisjes meededen aan het onderzoek dan is dat een flink karwei. In Excel kun je met behulp van **draaitabellen** snel dergelijke kruistabellen maken.
Een draaitabel is een tabel die data uit een nette dataset op een dynamische manier kan samenvatten, rangschikken en groeperen. Een draaitabel kan onder andere automatisch kruistabellen maken met aantallen, gemiddelden, standaardafwijkingen en minima.

[►EXCEL] Neem de Excelmodule **Draaitabellen** door.

Merk op dat een kruistabel met aantallen een frequentieverdeling is.

DEELNEMERS ONDERZOEK

	M	V	
Alabama	6	10	16
Alaska	0	1	1
Arizona	12	8	20
Arkansas	4	3	7
California	48	48	96
Colorado	11	12	23
⋮	⋮	⋮	⋮
Wisconsin	6	8	14
Wyoming	0	1	1
	402	396	798

10 Open het bestand JongerenSocialeNetwerken.xlsx.

a Maak een kruistabel met aantallen
- met als rij-variabele *Leeftijd* en kolomvariabele *Regio*
- met als rij-variabele *Regio* en kolomvariabele *Woonomgeving*.

b Maak kruistabellen waarmee de volgende vragen te beantwoorden zijn.
- Hoeveel van de onderzochte personen hebben zowel een mobiele telefoon als een tablet?
- Hoeveel jongens van 15 jaar waren bij dit onderzoek betrokken?

Door variabelen in rijen of kolommen te verwisselen (draaien) kun je de brongegevens op andere manieren samenvatten.

11 In het bestand Speerwerpen.xlsx is van elk van de 80 speerwerpers de beste afstand genoteerd die ze hebben geworpen bij de kwalificatieronde voor de Olympische Spelen van 2012.
Maak een draaitabel waarin de kleinste, de gemiddelde en de grootste geworpen afstanden zijn af te lezen, uitgesplitst naar geslacht en vervolgens naar groep. Rond de gemiddelden af op cm.

R 12 Ga uit van de dataset JongerenSocialeNetwerken.xlsx.
Maak een kruistabel met aantallen met de variabelen *DeeltFotosVanZichzelf* en *DeeltVideosVanZichzelf*.
Wat is de betekenis van het getal 12 in de tabel?

A 13 Ga uit van de dataset JongerenSocialeNetwerken.xlsx.
De variabele *FB_Vrienden* geeft aan hoeveel Facebookvrienden iemand heeft. De waarde #Null! geeft aan dat de vraag niet gesteld is aan deze persoon omdat deze geen Facebookaccount heeft.
Susanne beweert dat uit de dataset blijkt dat van de jongeren die een Facebookaccount hebben

- ze naarmate ze ouder zijn steeds meer Facebookvrienden hebben
- meisjes over het algemeen meer Facebookvrienden hebben dan jongens.

a Onderzoek of Susanne gelijk heeft. Licht je antwoord toe met behulp van één draaitabel met gemiddelde aantallen Facebookvrienden.

Michiel wil onderzoeken of de aantallen Facebookvrienden van jongeren op het platteland (Rural) afwijken van die van jongeren in de stad (Urban). Hij bedenkt dat ondervraagden zonder Facebookaccount nul Facebookvrienden hebben. Daarom hercodeert hij de variabele *FB_vrienden*.

b Maak voor Michiel de kruistabel die de gemiddelde aantallen facebookvrienden voor de genoemde woonomgevingen weergeeft uitgesplitst naar geslacht. Welke conclusie trekt Michiel?

O 14 Het bestand LeerlingGegevens.xlsx bevat informatie over leerlingen van het Playfair College.

a Maak met behulp van een draaitabel een frequentieverdeling van alle voorkomende lichaamslengtes. Waarom geeft deze frequentieverdeling geen goed beeld van de verdeling?

b Voeg aan de dataset de nieuwe variabele *LengteKlasse* toe, die aangeeft tot welk van de lengteklassen 150-159, 160-169, …, 180-189, > 189 de leerling behoort.

In de klasse 150-159 zitten alle waarnemingsgetallen vanaf 150 tot en met 159.

c Maak met behulp van een draaitabel een frequentieverdeling bij de variabele *LengteKlasse*.

d Welk type diagram zou je gebruiken om de frequentieverdeling van vraag c grafisch weer te geven?

Theorie B Groeperen en draaigrafieken

In opgave 14 heb je door middel van hercoderen een klassenindeling gemaakt.
Omdat het maken van klassenindelingen vaak voorkomt bij statistische onderzoeken is het mogelijk dit in Excel gemakkelijker en dynamischer te doen. Het maken van een klassenindeling komt neer op het **groeperen** van de waarden van een variabele. De waarden van de rijvariabele *Lengte* in de draaitabel van opgave 14a kun je groeperen door rechts te klikken op een van de waarden in de draaitabel en te kiezen voor "Groeperen...". Hoe dit in zijn werk gaat en hoe deze optie omgaat met gehele getallen en kommagetallen leer je in de volgende Excelmodule. In die module leer je ook hoe je bij een draaitabel een diagram maakt dat met de draaitabel mee verandert. Zo'n dynamisch diagram bij een draaitabel heet een **draaigrafiek**.

[►EXCEL] Neem de Excelmodule **Groeperen en draaigrafieken** door.

Afspraak
Maak in deze paragraaf de genoemde diagrammen met behulp van draaitabellen/draaigrafieken. Maak elke grafiek netjes op. Zorg dus voor voldoende informatie bij de assen, een passende grafiektitel en zo nodig een legenda.

15 Ga uit van de dataset StudentenEnquete.xlsx.

a Maak een cirkeldiagram bij de variabele *Groep*. Zorg ervoor dat in het diagram de aantallen studenten per groep te zien zijn.

b Maak een staafdiagram bij de variabele *Lengte*. Neem als klassen < 155, 155-159, …, 190-195, > 195.

c Maak een histogram bij de variabele *UrenSlaap*. Neem als klassen < 5, 5 −< 6, … 9 −< 10, ≥ 10.

d Maak een cirkeldiagram bij de variabele *AantalBroersEnZussen*. Meer dan 5 broers en zussen komt niet vaak voor. Groepeer daarom de aantallen broers en zussen boven 5 in een groep met de naam 'meer dan 5'.

Maak je een klassenindeling van kommagetallen dan noteert Excel de klasse 5 −< 6 als 5-6. De klasse ≥ 10 wordt dan genoteerd als > 10.

16 Het bestand Speerwerpen.xlsx bevat de resultaten van de kwalificatieronde speerwerpen voor de Olympische Spelen van 2012.

a Maak een draaitabel met een frequentieverdeling van de geworpen afstanden en groepeer de geworpen afstanden in de klassen < 55, $55 -< 60$, …, $85 -< 90$.

b Maak bij de tabel van vraag a een histogram.

c Met wat voor soort verdeling heb je hier te maken?

d Gebruik bij het histogram de variabele *Geslacht* als legenda-variabele. Hoe is het antwoord op vraag c te verklaren?

In de Excelmodule Groeperen en draaigrafieken heb je geleerd wat een legenda-variabele is.

Bij zowel de mannen als de vrouwen waren de atleten opgedeeld in de groepen A en B.

e Pas het histogram van vraag d aan zodat zowel bij de mannen als bij de vrouwen de resultaten van de A- en de B-groep snel vergeleken kunnen worden.

f Volgens een sportjournalist kun je beter in groep A zitten dan in groep B, want "de atleten in groep A gooiden verder". Geef commentaar.

17 Ga uit van de dataset Facebook.xlsx. Deze dataset is het resultaat van een enquête onder Facebookgebruikers in de VS.

a Maak een stapeldiagram waarin is af te lezen hoeveel jongens en hoeveel meisjes er van elke leeftijd zijn ondervraagd.

b Met een cirkeldiagram kan weergegeven worden welk percentage van deze Facebookgebruikers twittert en welk percentage niet.
Maak dit cirkeldiagram en zorg ervoor dat de percentages in het diagram zichtbaar zijn.

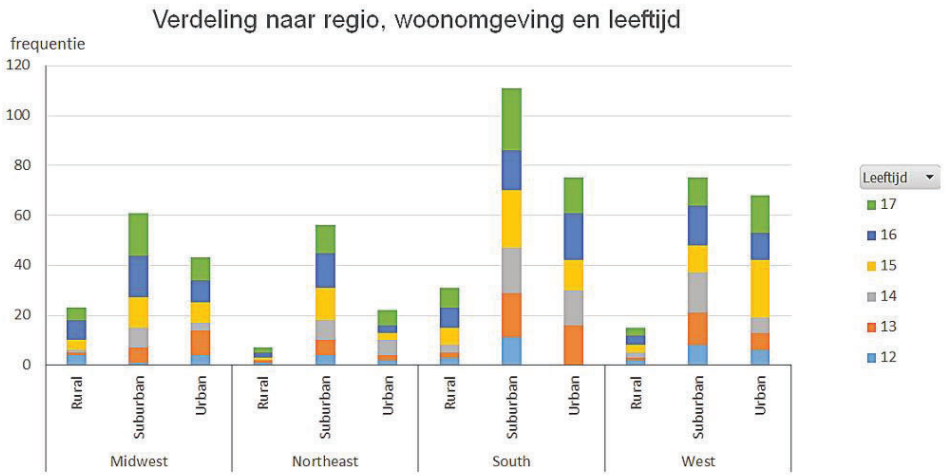
Simone heeft voor een praktische opdracht deze dataset bestudeerd. Volgens haar hebben meisjes gemiddeld meer Facebookvrienden (*FB_Vrienden*) dan jongens. Om deze bewering te onderbouwen wil ze in haar verslag een staafdiagram opnemen waarin het gemiddeld aantal Facebookvrienden voor jongens en voor meisjes is te zien. Ook wil ze zowel voor de jongens als voor de meisjes onderzoeken of er een verband is tussen de leeftijd en het aantal Facebookvrienden.

c Maak voor Simone het genoemde staafdiagram en breid het staafdiagram uit zodat Simone haar vervolgonderzoek kan doen.

d Welke conclusie trekt Simone?

18 Ga uit van de dataset Facebook.xlsx.

a Maak bij deze dataset de draaigrafiek van figuur 8.8 die informatie geeft over de leeftijd, regio en leefomgeving van de deelnemers aan de enquête.



figuur 8.8

b Maak een stapeldiagram dat de verdeling over de regio's van de mannelijke deelnemers van 15 jaar weergeeft.

c Maak een stapeldiagram dat per leeftijd de procentuele verdeling van de deelnemers naar woonomgeving weergeeft.

A 19 Ga uit van de dataset JongerenSocialeNetwerken.xlsx.
De variabelen *DeeltFotosVanZichzelf* en *DeeltVideosVanZichzelf* hebben de waarde !null gekregen als de deelnemer deze vraag niet heeft beantwoord, omdat hij of zij geen sociale netwerken gebruikt.

a Maak een draaigrafiek van het type 100% gestapelde kolom. Zorg ervoor dat in het diagram voor elke leeftijd, uitgesplitst naar geslacht te zien is hoe de waarden van de variabele *DeeltFotosVanZichzelf* zich verhouden.

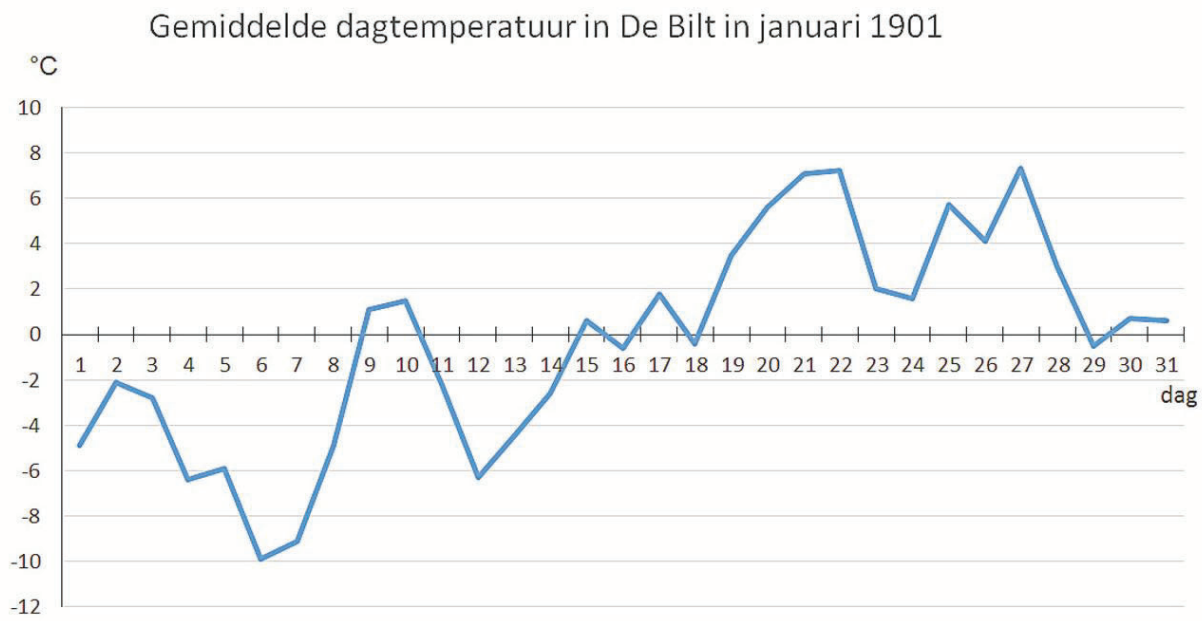
Niels en Dayal bekijken de draaitabel van vraag a. Volgens Niels deelt ongeveer 22% van de 16-jarige meisjes uit het onderzoek geen foto's van zichzelf op sociale netwerksites. Volgens Dayal is dat ongeveer 26%.

b Hoe komt Niels op ongeveer 22% en hoe komt Dayal op ongeveer 26%? Wie heeft gelijk? Licht toe.

c Hercodeer de variabele *DeeltFotosVanZichzelf* zodat alleen nog de waarden ja en nee voorkomen en maak opnieuw de draaigrafiek van vraag a.

d Voeg aan de draaitabel van vraag c een variabele toe zodat in het stapeldiagram de verhoudingen in beeld worden gebracht tussen de groepen die van zichzelf
- zowel foto's als video's delen
- alleen foto's delen
- alleen video's delen
- geen van beide delen.

O20 Het bestand TemperatuurHistorie.xlsx bevat een dataset met alle door het weerstation in De Bilt gemeten dagtemperaturen in de periode van 1 januari 1901 tot en met 14 augustus 2014.



figuur 8.9

a Maak het lijndiagram van figuur 8.9. Ga als volgt te werk.
- Selecteer de temperaturen van januari 1901.
- Voeg het lijndiagram in via Invoegen > Grafieken.
- Maak het lijndiagram netjes op.

Nathalie heeft een artikel gelezen over de opwarming van de aarde. Ook in Nederland zou een stijgende trend te zien zijn in de gemiddelde dagtemperaturen. Nathalie wil deze stijgende trend in beeld brengen door alle temperaturen in de dataset te selecteren en er een lijndiagram bij te maken.

b Vind je dat in dit lijndiagram de stijgende trend goed in beeld wordt gebracht? Licht toe.

8

Theorie C Groeperen van datums

Het lijndiagram van opgave 20b geeft geen goed beeld van de stijgende trend in temperaturen. Dit komt doordat het lijndiagram uit veel te veel gegevenspunten bestaat. In het diagram is het patroon van de jaargetijden nog enigszins te herkennen, maar deze verbergt grotendeels de trend. Het zou beter zijn een lijndiagram te maken met als gegevenspunten de gemiddelde jaartemperaturen of bijvoorbeeld met de gemiddelde maandtemperaturen van de maand mei. Hoe je deze diagrammen maakt leer je in de volgende Excelmodule.

[►EXCEL] Neem de Excelmodule **Groeperen van datums** door.

21 In het bestand WoningVerkoop.xlsx zie je informatie over woningverkopen in Nederland in de periode januari 1995 - juni 2014. De variabele *AantalVerkochteWoningen* geeft aan hoeveel woningen er in totaal in de betreffende maand zijn verkocht.

a Groepeer de variabele *Periode* op jaren én kwartalen en maak voor de periode 1995-2007 een lijndiagram van de verkochte aantallen woningen. Welk patroon is in het diagram te herkennen?

b Maak met behulp van een draaigrafiek een lijndiagram waarin voor de jaren 1995-2013 is te zien hoeveel woningen er dat jaar verkocht zijn en wat de gemiddelde verkoopprijs van een woning was.

c Bij de analyse van de woningmarkt wordt vaak gekeken naar de totale woningwaarde van de verkochte woningen. De totale woningwaarde is het aantal verkochte woningen vermenigvuldigd met de gemiddelde verkoopprijs.
Voeg aan het lijndiagram van vraag b een grafiek toe van de totale woningwaarde in honderdduizenden euro's.

A22 Regelmatig verschijnen er in de krant en op internet artikelen over weerrecords en hoeveel het weer afwijkt van normaal. Met het begrip normaal wordt tot en met 2020 het gemiddelde weer van de jaren 1981-2010 bedoeld. Dit wordt ook wel het langjarig gemiddelde genoemd. Het artikel hieronder gaat over het weer in De Bilt in juli 2014. Op enkele plaatsen ontbreken de getallen.

Juli 2014: Zeer warm, vrij nat en vrijwel normale hoeveelheid zon

Met een gemiddelde temperatuur van circa … °C tegen een langjarig gemiddelde van … °C was juli een zeer warme maand. Juli 2014 eindigt daarmee op de ...e plaats in de lijst van warmste julimaanden sinds 1951.
Het is ook de achtste maand op rij die warmer dan normaal is verlopen. De maand begon echter vrij koel, toen met een noordelijke stroming juist koele lucht werd aangevoerd. Daarna werd het warmer. Echte hitte volgde halverwege de maand.
Het aantal uren zonneschijn per dag week niet veel af van het langjarig gemiddelde en komt in De Bilt gemiddeld uit op … uren, tegen … uren normaal. In De Bilt werd … mm neerslag geregistreerd tegen een langjarig gemiddelde van … mm.

Ga uit van de dataset WeerDeBilt.xlsx. Op het tabblad Historie zijn de gegevens vanaf 1951 tot en met 14-08-2014 verzameld. Op het tabblad Tijdvak 1981-2010 zijn de gegevens van de jaren 1981-2010 nog eens apart opgenomen en op het tabblad Juli 2014 zijn de gegevens van juli 2014 apart opgenomen.

a Onderzoek welke getallen in de eerste alinea van het artikel op de puntjes moeten staan. Doe dit ook voor de laatste alinea.

b Maak op een leeg tabblad een nieuwe tabel waarin je zowel de gemiddelde dagtemperaturen van juli 2014 als die van juli in de periode 1981-2010 verzamelt. Maak bij deze tabel een lijndiagram.

c Maak een diagram zoals bij vraag b, maar nu voor de hoeveelheid neerslag. Geef een beschrijving van de neerslag in juli 2014 in de Bilt ten opzichte van wat normaal is in juli.

8.3 Data analyseren

O23 Zie het artikel hieronder.

10 procent Nederlanders gelooft in 'superfood'

Tien procent van de Nederlanders gelooft dat het noodzakelijk is naast hun gewone voeding zogenoemde superfoods tot zich te nemen. Deze mensen vertrouwen er bijvoorbeeld op dat gojibessen de kans op kanker verminderen of dat chiazaad het verouderingsproces vertraagt, concludeert het Voedingscentrum uit een steekproef die het heeft laten uitvoeren onder duizend mensen.

Geef het 95%-betrouwbaarheidsinterval voor het percentage van de Nederlanders dat gelooft in superfood. Neem aan dat 100 respondenten aangaven te geloven in superfood.

O24 De afgelopen jaren konden de werknemers van de firma SAMX kiezen uit vier typen kerstpakketten. Nadat elke werknemer zijn keuze had doorgegeven werden de verschillende pakketten in de juiste aantallen besteld. Ook dit jaar geeft elke werknemer weer zijn keuze door, maar omdat SAMX kosten kan besparen door nog maar één type pakket te bestellen, wordt voor alle werknemers het type besteld dat het meest gekozen is. Werknemer Floris probeert erachter te komen welk type kerstpakket hij zal ontvangen, door willekeurig 500 collega's te vragen naar hun keuze. In het bestand Kerstpakketten.xlsx zie je het resultaat van zo'n steekproef van lengte 500. In de cellen G5 t/m G8 zie je van elk van de pakketten de steekproefproportie.
Met de functietoets F9 wordt steeds een nieuwe steekproef genomen. Dit gebeurt ook als je een nieuwe formule invoert.

a Bereken in de cellen H5 t/m H8 de standaardafwijkingen bij de steekproefproporties en bereken van de bijbehorende 95%-betrouwbaarheidsintervallen de ondergrenzen in de cellen I5 t/m I8 en de bovengrenzen in J5 t/m J8.

b Denk je dat Floris een goede voorspelling kan maken van het kerstpakket dat hij dit jaar krijgt?

95%-betrouwbaarheids-interval $[\hat{p} - 2\sigma, \hat{p} + 2\sigma]$ met

$\hat{p}$ = steekproefproportie en

$\sigma = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

Een wortel bereken je in Excel met de functie WORTEL.

Theorie A Populatiekenmerken

Bij veel onderzoeken wordt een steekproef genomen met als doel een indicatie te krijgen van kenmerken van de populatie. Je weet al dat je een goede schatting van een populatieproportie kunt krijgen door van een representatieve steekproef de steekproefproportie te berekenen. Immers bij een representatieve steekproef zal p ongeveer gelijk zijn aan $\hat{p}$. Je weet ook dat in plaats van één waarde voor de geschatte proportie vaak een betrouwbaarheidsinterval wordt gegeven.
Bij nominale en ordinale variabelen kun je van elk van de voorkomende waarden de proportie berekenen.
Bij variabelen gemeten op interval- of rationiveau krijg je een indicatie van het populatiegemiddelde aan de hand van het steekproefgemiddelde. Ook daarbij zijn betrouwbaarheidsintervallen te berekenen.

Voor het berekenen van een betrouwbaarheidsinterval voor een gemiddelde gebruik je in Excel de functie
BETROUWBAARHEID(*alfa*, *standaarddev*, *grootte*).
Bij deze functie

- is *grootte* de steekproefomvang
- is *standaarddev* de steekproefstandaardafwijking
- geeft *alfa* het betrouwbaarheidsniveau aan.
  Bij een betrouwbaarheidsniveau van 95% hoort *alfa* = 0,05 en bij een betrouwbaarheidsniveau van 80% hoort *alfa* = 0,20.

Bij een steekproef met steekproefgemiddelde $\bar{X}$ is het betrouwbaarheidsinterval
[$\bar{X}$ − BETROUWBAARHEID, $\bar{X}$ + BETROUWBAARHEID].

De functie BETROUWBAARHEID is niet te gebruiken bij het berekenen van een betrouwbaarheidsinterval voor een proportie.

8

25 Het bestand Wonen.xlsx bevat een deel van de dataset uit het Woononderzoek Nederland van het CBS. Het onderzoek richt zich op alle huishoudens die op 1 januari 2012 huisvesting in Nederland hadden.
Ga uit van dit bestand en bereken

a de steekproefproportie van het aantal huishoudens dat in een huurwoning woont

b het 95%-betrouwbaarheidsinterval van de proportie huishoudens die in een hoekwoning woont

c van de personen die in een koopwoning wonen het 95%-betrouwbaarheidsinterval van de proportie die in een vrijstaand huis woont.

26 Ga uit van het bestand Wonen.xlsx. De WOZ-waarde van een huis is de waarde zoals de gemeente die heeft vastgesteld. Diverse belastingen en heffingen worden gebaseerd op basis van deze Waardering Onroerende Zaken.
Bereken in duizenden euro's nauwkeurig

a het 95%-betrouwbaarheidsinterval van de gemiddelde WOZ-waarde van een woning in Nederland

b het 95%-betrouwbaarheidsinterval van de gemiddelde WOZ-waarde van vrijstaande koopwoningen

c het 80%-betrouwbaarheidsinterval van de gemiddelde WOZ-waarde van woningen waarvan de bewoners vaak overlast hebben van zowel het verkeer als van jongeren.

A27 Ga uit van het bestand Wonen.xlsx.

a Bereken het 68%-betrouwbaarheidsinterval van de proportie huishoudens die tevreden is over de woning waarin ze wonen.

b Bereken in duizenden euro's nauwkeurig het 90%-betrouwbaarheidsinterval van de gemiddelde WOZ-waarde van woningen die bewoond worden door huishoudens die zeer tevreden zijn over de woning waarin ze wonen.

O28 Ga uit van het bestand Wonen.xlsx.

a Maak bij de variabele *WozWaarde* een frequentieverdeling voor huurwoningen en voor koopwoningen en maak hierbij een lijndiagram. Gebruik voor de WOZ-waarden de klassen $< 100\,000$, 100 000-199 999, 200 000-299 999, …, $> 700\,000$.
Noem een overeenkomst en een verschil in de WOZ-waarden van huurwoningen en koopwoningen.

b De gemiddelde WOZ-waarde van koopwoningen verschilt van die van huurwoningen.
Bereken dit verschil. Zou je dit verschil klein, middelmatig of groot noemen?

c Zie de tabel hiernaast.
Welk verschil merk je op? Zou je dit verschil klein, middelmatig of groot noemen?

d Zie de tabel hiernaast.
Omschrijf het verschil in tevredenheid over de woning tussen huishoudens in een huurwoning en in een koopwoning.
Zou je dit verschil klein, middelmatig of groot noemen?

e De variabele *Woning* is een nominale variabele.
Geef van elk van de drie andere variabelen die in deze opgave voorkomen het meetniveau.

AANTALLEN WONINGEN

	huurwoning	koopwoning
eengezinswoning	20 002	62 382
meergezinswoning	23 658	11 016

AANTALLEN WONINGEN

tevredenheid woning	huurwoning	koopwoning
1. zeer ontevreden	822	130
2. ontevreden	2090	626
3. niet tevreden, maar ook niet ontevreden	5636	2884
4. tevreden	22 216	29 760
5. zeer tevreden	12 896	39 998
	43 660	73 398

Theorie B Groepen vergelijken

In opgave 28 heb je gekeken naar verschillen tussen huur- en koopwoningen ten aanzien van de WOZ-waarde, de woningvorm en de mate van tevredenheid over de woning. Het omschrijven van het verschil is vaak eenvoudiger dan aangeven hoe groot het verschil is. Bij vraag b heb je het verschil in WOZ-waarde weergegeven met één getal. Het weergeven van een verschil met een getal wordt **kwantificeren van een verschil** genoemd. Het kwantificeren van het verschil tussen huur- en koopwoningen ten aanzien van de WOZ-waarde is eenvoudig, omdat de variabele *WozWaarde* een ratiovariabele is. Ook bij nominale en ordinale variabelen zijn verschillen te kwantificeren.

Verschillen kwantificeren bij nominale variabelen

De kruistabel hiernaast is gemaakt om te onderzoeken hoe groot het verschil is in profielkeuze tussen havo- en vwo-leerlingen.

	profiel		
	maatschappij	natuur	
havo	238	94	332
vwo	107	124	231
	345	218	563

- *Percentageverschil (PV)*
  Van de havoleerlingen heeft $\frac{238}{332} \times 100\% \approx 71{,}7\%$ een maatschappijprofiel.

  Bij de vwo'ers is dit $\frac{107}{231} \times 100\% \approx 46{,}3\%$.

  Het percentageverschil *PV* is $71{,}7\% - 46{,}3\% = 25{,}4\%$.
- *Odds-ratio (OR)*

  $$\text{de odds-ratio} = \frac{\text{grootste kruisproduct}}{\text{kleinste kruisproduct}}$$

  De kruisproducten zijn $238 \cdot 124 = 29\,512$ en $107 \cdot 94 = 10\,058$, dus

  $$OR = \frac{29512}{10058} \approx 2{,}9.$$

  Dit wil ruwweg zeggen dat een havoleerling 2,9 keer zo vaak kiest voor een maatschappijprofiel als een vwo-leerling en ook dat een vwo-leerling 2,9 keer zo vaak voor een natuurprofiel kiest als een havoleerling.

De vuistregels hieronder worden gebruikt om te bepalen of het verschil klein, middelmatig of groot genoemd wordt.

PV	***OR***	**het verschil is**
$PV \leq 15\%$	$OR < 2$	klein
$15\% < PV \leq 30\%$	$2 \leq OR \leq 3$	middelmatig
$PV > 30\%$	$OR > 3$	groot

Je ziet dat, zowel op basis van het *PV* als op basis van de *OR*, het verschil in profielsoort tussen havo-leerlingen en vwo-leerlingen middelmatig is. Het is niet altijd zo dat het *PV* en de *OR* dezelfde aanduiding van de grootte van het verschil geven.

Verschillen kwantificeren bij ordinale variabelen

Maximale cumulatieve percentageverschil (max.Vcp)

Deelnemers aan een onderzoek naar de factoren die van invloed zijn op de profielkeuze hebben op een schaal van 1 tot en met 5 aangegeven in hoeverre de beoogde vervolgopleiding bij de keuze een rol speelde. De resultaten zie je in de tabel hiernaast.

	maatschappij	natuur
1	25	17
2	29	9
3	78	30
4	96	64
5	95	80
	323	200

Om het verschil tussen leerlingen die een maatschappijprofiel hebben gekozen en die een natuurprofiel hebben gekozen te kwantificeren, kun je het maximale verschil van de cumulatieve percentages (*max.Vcp*) gebruiken.

In de tabel hieronder zijn de cumulatieve percentages en de verschillen van deze percentages berekend.

	maatschappij	natuur	*Vcp*
1	7,7%	8,5%	0,8%
2	16,7%	13%	3,7%
3	40,8%	28%	12,8% ← *max.Vcp*
4	70,7%	60%	10,7%
5	100%	100%	0%

$\frac{25}{323} \times 100\% \approx 7{,}7\%$

$\frac{25 + 29}{323} \times 100\% \approx 16{,}7\%$

...

Je ziet dat het maximale verschil van de cumulatieve percentages 12,8% is, dat wil zeggen dat het verschil tussen maatschappij- en natuurprofielen klein is.

De vuistregels hieronder worden gebruikt om te bepalen of het verschil klein, middelmatig of groot genoemd wordt.

max.Vcp	**het verschil is**
$max.Vcp \leq 15\%$	klein
$15\% < max.Vcp \leq 30\%$	middelmatig
$max.Vcp > 30\%$	groot

8

29 Bereken bij de tabel van opgave 28c zowel het *PV* als de *OR* en geef aan of dit op een klein, middelmatig of groot verschil wijst.

30 Ga uit van het bestand Wonen.xlsx.
Maak met behulp van een draaitabel de frequentieverdeling van opgave 28d. Geef de waarden weer als relatieve cumulatieve frequenties.
Bereken het *max.Vcp*. Is het verschil klein, middelmatig of groot?

Relatieve cumulatieve frequenties in een draaitabel: Rechtsklik op een waarde in de draaitabel. Kies **waarden weergeven als** en vervolgens **% voorlopig totaal in**.

31 Ga uit van het bestand Wonen.xlsx.
Omschrijf het verschil in besteedbaar inkomen van huishoudens die in een koophuis wonen ten opzichte van huishoudens die in een huurhuis wonen. Illustreer dit verschil met een geschikt diagram. Kwantificeer het verschil en geef aan of dit verschil klein, middelmatig of groot is.

32 Ga uit van het bestand Wonen.xlsx.
Er bestaat een verschil in tevredenheid over de woonomgeving tussen huishoudens die wonen in een vrijstaande woning en huishoudens die wonen in een niet-vrijstaande woning. Omschrijf dit verschil en geef aan of het verschil klein, middelmatig of groot is.

33 Een politicus beweert dat verkeersoverlast en overlast van jongeren hand in hand gaat.
Maak een kruistabel met de variabelen *OverlastJongeren* en *OverlastVerkeer* uit het bestand Wonen.xlsx. Maak in de tabel alleen onderscheid tussen de waarden vaak en niet-vaak. Bereken de *OR* en geef hiervan een interpretatie.
Ben je het met de politicus eens?

34 De variabele *Plaats* in de dataset Wonen.xlsx geeft informatie over de woonplaats van de ondervraagden. G4 geeft aan dat de ondervraagde woont in een van de vier grote steden (Amsterdam, Rotterdam, Den Haag, Utrecht), onder G27 vallen Groningen, Leeuwarden, Emmen, Almelo, Deventer, Enschede, Hengelo, Zwolle, Arnhem, Nijmegen, Amersfoort, Lelystad, Alkmaar, Haarlem, Zaanstad, Dordrecht, Leiden, Schiedam, Breda, Eindhoven, Helmond, Den Bosch, Tilburg, Heerlen, Maastricht, Venlo en Sittard-Geleen. G4 en G27 samen noemen we G31.
Omschrijf het verschil in aanwezigheid van bekladding, rommel en hondenpoep tussen G31-steden en overige woonplaatsen. Geef steeds aan of het verschil klein, middelmatig of groot is.
Welk verschil is het grootst?

8.4 Onderzoeken

Theorie A Statistisch onderzoek

Bij een statistisch onderzoek probeert men antwoord op een vraag te vinden door het verzamelen, verwerken en interpreteren van gegevens.

De **statistiek** houdt zich bezig met methoden en technieken voor het verzamelen van gegevens en ook voor het verwerken, analyseren en presenteren daarvan.
In een deugdelijk statistisch onderzoek zijn in ieder geval de volgende fasen te onderscheiden.
1 het onderzoeksontwerp
2 het verzamelen van gegevens
3 het verwerken en analyseren van de verzamelde gegevens
4 het trekken van conclusies

In de vorige paragrafen had je steeds te maken met de laatste twee fasen. In deze paragraaf zijn ook de eerste twee fasen van belang.

Het onderzoeksontwerp
Het **onderzoeksontwerp** omvat een *onderzoeksonderwerp*, een *probleemstelling* en de *hoofdvraag* die daaruit volgt. Bij de hoofdvraag worden vaak *deelvragen* geformuleerd die helpen een antwoord te vinden op de hoofdvraag. In de fase van het onderzoeksontwerp wordt er nagedacht over welke variabelen een rol spelen bij de hoofd- en deelvragen en hoe de gegevens verzameld moeten worden. Bovendien is er aandacht voor de relevantie van het onderzoek en de haalbaarheid. In de onderzoeksfase moet worden bepaald op welke populatie het onderzoek zich richt.

35 Bij een onderzoek probeert men een antwoord te vinden op de vraag "Wat valt er te zeggen over de tijdsbesteding van de leerlingen van het Staring College?".
Hierbij worden onder andere de volgende deelvragen gesteld.
1 Besteden leerlingen die veel uren werken voor een bijbaan minder tijd aan huiswerk?
2 Besteden de bovenbouwleerlingen van het vwo meer tijd aan hun huiswerk dan de onderbouwleerlingen van het vwo?
3 Kijken leerlingen die veel sporten minder tv dan leerlingen die weinig sporten?
4 Sporten jongens meer dan meisjes?
a Bedenk nog vier deelvragen die passen bij de hoofdvraag van dit onderzoek.
b Geef voor elk van de gegeven deelvragen aan welke variabelen hierbij een rol spelen. Doe dit ook voor de deelvragen die je zelf hebt bedacht bij vraag a.

36 Er wordt een onderzoek gedaan naar de factoren die van invloed zijn op de profielkeuze van de derdeklasleerlingen. Bedenk vier deelvragen bij dit onderzoek en geef bij elke deelvraag een variabele aan die erbij hoort.

Theorie B Gegevens verzamelen

Om antwoorden te vinden op de onderzoeksvragen verzamel je gegevens. Er zijn verschillende manieren om aan gegevens te komen.

Bestaand cijfermateriaal

Voor veel onderzoeken kan gebruik worden gemaakt van bestaand cijfermateriaal. Op internet zijn veel gegevens te vinden. Bekende bronnen voor gegevens over Nederland en haar inwoners zijn de site van het Centraal Bureau voor de Statistiek (CBS) en het Open data-portal van de Nederlandse overheid.

De datasets die je tot nu toe gebruikt hebt zijn zo bewerkt dat ze te begrijpen zijn zonder veel nadere toelichting. De meeste datasets zijn echter niet te lezen zonder codeboek. In een **codeboek** wordt aangegeven hoe de onderzoeksgegevens gecodeerd zijn. Het codeboek geeft voor elke variabele aan wat de betekenis van de variabele is en welke enquêtevraag of waarneming erbij hoort. Het geeft ook aan welke codes gebruikt zijn voor de antwoorden en welke eenheid erbij hoort.
Hieronder zie je een deel van een dataset van het CBS en het bijbehorende codeboek.

In een codeboek wordt aangegeven hoe de onderzoeksgegevens gecodeerd zijn.

GESLACHT	FAMICONT	UULiSP
1	1	1
1	1	3
1	5	3
2	1	4
1	1	3
2	1	5

GESLACHT Geslacht
1 Man, 2 Vrouw

FAMICONT Contact met familieleden
1 Minstens 1 keer per week
2 Vaker dan 1 keer per maand, maar niet wekelijks
3 1 keer per maand
4 Minder dan 1 keer per maand
5 Zelden of nooit

UULiSP Aantal uren dat de respondent per week sport

Gegevens verwerven

Bij veel onderzoeken zullen de juiste gegevens echter niet beschikbaar zijn. De onderzoeker zal dan door middel van een enquête of door bijvoorbeeld een simulatie of een experiment de gegevens moeten verwerven.

Een enquête is een veelgebruikte methode om gegevens te verzamelen. Het opstellen van een goede vragenlijst is echter nog niet zo gemakkelijk. Hiernaast zie je een aantal tips van het CBS voor het maken en afnemen van een enquête. De deelnemers aan de enquête worden de **respondenten** genoemd.

Tip 1: Geef duidelijk aan over welke periode de vraag gaat.
Tip 2: Stel één vraag per vraag.
Tip 3: Maak korte vragen en zinnen.
Tip 4: Geef niet je eigen mening in vragen.
Tip 5: Geef een keuze uit antwoorden.
Tip 6: Zorg dat de antwoordmogelijkheden elkaar niet overlappen.
Tip 7: Zoek je respondenten op een logische plek.
Tip 8: Maak je vragenlijst niet te lang!

37 Lever commentaar op de volgende enquêtevragen.

a U voelt zich zeker ook steeds onveiliger in deze buurt?

b Eet u ook steeds vaker vegetarisch en vindt u ook dat u daarom gezonder leeft?

c Doet u veel aan sport?

d Hoe vaak heeft u vorige week de website Nu.nl bezocht?
Er kan gekozen worden uit de antwoorden
I elke dag II nooit III 1 tot 3 keer IV 3 tot 5 keer.

38 Stel bij elk van de enquêtevragen van opgave 37 een nieuwe vraag op waarbij je rekening houdt met alle tips van het CBS.

39 Stel bij elk van de vier gegeven deelvragen van opgave 35 een geschikte enquêtevraag op.

Onderzoek

Deze onderzoeksopdrachten kun je verwerken in een verslag of presentatie.

I Top 2000

De Top 2000 is een jaarlijks radioprogramma van het Nederlandse station NPO Radio 2. Het wordt sinds 1999 uitgezonden aan het eind van elk kalenderjaar. In het programma worden de 2000 populairste platen aller tijden gedraaid, zoals die in de weken voorafgaand aan de uitzending door middel van een stemming via internet door de luisteraars zijn bepaald.
In het bestand 15jaarTop2000.xlsx zijn gegevens verzameld van de Top 2000 van de jaren 1999 tot en met 2013.

- Zoek op internet de gegevens van de laatste paar jaren en vul de dataset aan.
- Kies een onderzoeksonderwerp bij deze dataset. Formuleer hierbij een hoofdvraag en een aantal deelvragen. Analyseer de dataset en maak diagrammen.
- Schrijf een artikel voor een muziekblad over de Top 2000 waarin je de resultaten van je onderzoek verwerkt. Verwerk bovendien een aantal 'weetjes' in het artikel, bijvoorbeeld hoeveel dagen, uren en minuten er nodig zijn om alle liedjes van de top 2000 van 2013 non-stop achter elkaar te draaien.

II Krantenartikel

- Zoek op internet een dataset over een onderwerp naar keuze. Analyseer de gegevens en maak enkele relevante diagrammen. Trek conclusies en schrijf vervolgens een krantenartikel waarin je je bevindingen presenteert. Vermeld in het artikel de bron van de dataset.
- Voer een klein onderzoek uit onder je medeleerlingen of in de buurt waar je woont. Verwerf je gegevens door middel van een zelf opgestelde enquête. Zorg voor minimaal 40 respondenten. Analyseer de gegevens en maak enkele relevante diagrammen. Trek enkele conclusies en schrijf vervolgens een artikel voor de schoolkrant waarin je je bevindingen presenteert.

III Onderzoek controleren

Zoek, bijvoorbeeld op internet, een bestaande dataset waarbij een onderzoeksrapport of een artikel is gepubliceerd.
In het onderzoeksrapport of het artikel worden conclusies gepresenteerd, meestal ondersteund door enkele diagrammen, die op basis van de dataset zijn getrokken. In deze opdracht is het de bedoeling dat je die conclusies en de gepresenteerde diagrammen controleert op juistheid.
Verwerk je bevindingen in een kort verslag.

IV Zelf onderzoek doen

In deze opdracht ga je zelf een onderzoek doen, waarbij je de vier fasen van statistisch onderzoek zelf doorloopt. Je begint met het onderzoeksontwerp. Kies dus een onderwerp en draag een probleemstelling aan. Formuleer een hoofdvraag en deelvragen. Bedenk op wie het onderzoek betrekking heeft en hoe je je gegevens verzameld. Houd hierbij rekening met de haalbaarheid. Presenteer je onderzoeksontwerp bij je docent voor je verder gaat met de volgende fasen.
Verwerk je onderzoek in een onderzoeksrapport.

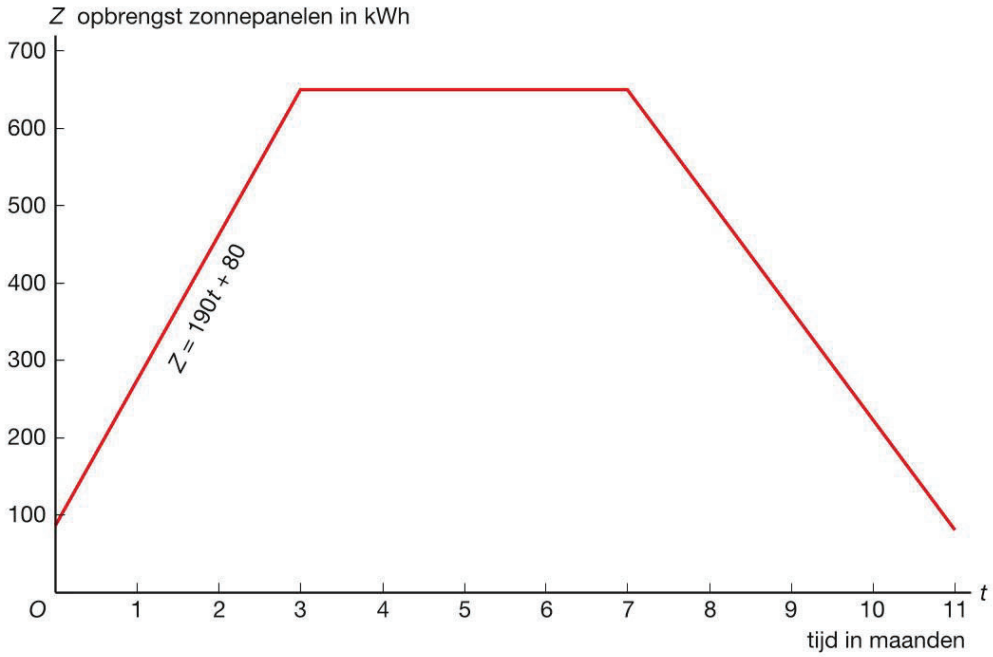
Fasen statistisch onderzoek

1. Onderzoeksontwerp
2. Gegevens verzamelen
3. Verwerken en analyseren
4. Conclusies trekken

Onderzoeksopgaven

Zonnepanelen

De familie Dijk heeft sinds enkele jaren zonnepanelen op het dak.
Meneer Dijk heeft een model gemaakt van de maandelijkse opbrengst in kWh van de panelen.
Volgens meneer Dijk is de opbrengst in januari 80 kWh en neemt dit lineair toe tot 650 kWh in april. Ook in de maanden mei tot en met augustus is de opbrengst 650 kWh. Daarna neemt de opbrengst lineair af tot 70 kWh in december.
Zie de grafiek in figuur O.1.



figuur O.1

Volgens mevrouw Dijk kloppen de opbrengsten van januari en december wel, maar is de maximale opbrengst van april tot en met augustus niet 650 kWh, maar 590 kWh.
Ook mevrouw Dijk gaat verder uit van lineaire verbanden.

Onderzoek hoeveel de jaarlijkse opbrengst volgens mevrouw Dijk minder is dan de jaarlijkse opbrengst volgens meneer Dijk.

Wit asfalt

Lees onderstaand krantenartikel.

Wit asfalt snelweg enorme besparing

Wegbeheerders kunnen tientallen miljoenen euro's besparen door lichtgekleurd asfalt te gebruiken. Daardoor is er veel minder verlichting nodig op snelwegen; en zelfs als er geen lampen zijn helpt het weggebruikers in het donker de weg te vinden.

Dura Vermeer

"De resultaten zijn spectaculair, want uit het verlichtingsonderzoek blijkt dat een vermindering van de openbare verlichting van wel 50% mogelijk is. Dit zou dus een enorme energiebesparing en daarmee samenhangende vermindering van de CO_2-emissie kunnen betekenen als dit toegepast wordt", zegt innovatiemanager Robbert Naus.

Vooralsnog is het nog wel een beetje duurder dan gewoon asfalt. "Dan hebben we het over drie euro per vierkante meter, oftewel 6,5% meer dan het oorspronkelijke asfalt. Dat betaalt zich binnen een jaar of tien echter makkelijk terug omdat er minder licht aan hoeft. En als het dan ook nog de kans op ongelukken vermindert en CO_2-vriendelijk is, betekent dit een win-winsituatie."

Een wegdek met een lengte van 5,3 kilometer en een totale breedte van 9 meter moet worden voorzien van nieuw asfalt. Langs het wegdek staan in totaal 298 lantaarns, die 3,3 cent per branduur per lantaarn kosten. De lantaarns branden op dit moment gemiddeld 9,5 uur per dag. Ga ervan uit dat door het gebruik van het duurdere, witte asfalt de kosten van de verlichting met 50% zullen afnemen.

Hoeveel jaar duurt het totdat de hogere kosten van het asfalt terugverdiend zijn door de lagere kosten van de verlichting? Rond je antwoord af op een geheel aantal jaren.

Woningwaardeontwikkeling

In de statistiek werkt men vaak met indexcijfers. Eén van de waarnemingsgetallen stel je 100. Daarna druk je de andere waarnemingsgetallen daarin uit.
Ken je in tabel I aan het jaar 2012 het indexcijfer 100 toe, dan is het indexcijfer van 2014 gelijk aan

$$\frac{16\,527}{23\,379} \cdot 100 \approx 70{,}7.$$

Weet je bij tabel II dat bij indexcijfer 115,2 het aantal 15 873 hoort dan is het aantal dat bij indexcijfer 72,6 hoort gelijk aan $\frac{15\,873}{115{,}2} \cdot 72{,}6 \approx 10\,003$.

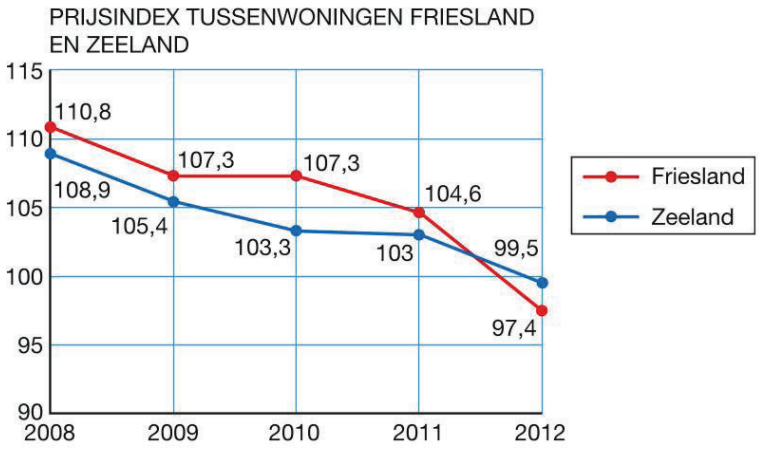
TABEL I

jaar	2012	2014
index	100	
aantal	23 379	16 527

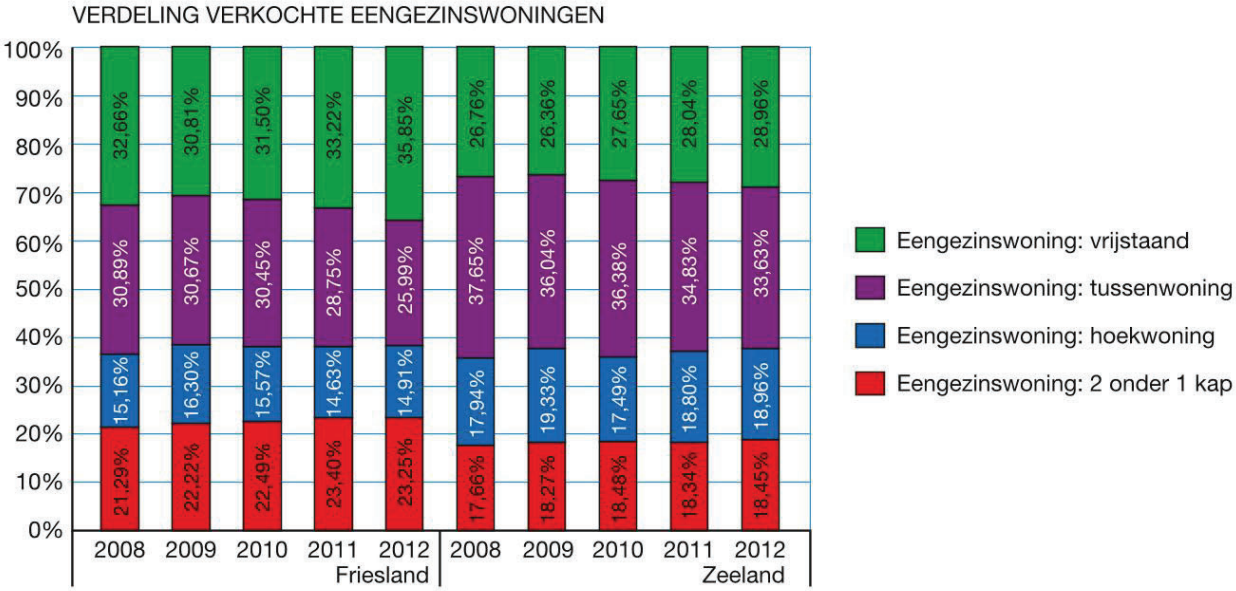
TABEL II

jaar	2011	2015
index	115,2	72,6
aantal	15 873	

De PBK (Prijsindex Bestaande Koopwoningen) geeft met indexcijfers aan hoe de marktwaarde van bestaande koopwoningen zich ontwikkelt met de jaren. Behalve deze index zijn er ook prijsindexen per type woning. In figuur O.2 zie je de indexen van tussenwoningen in Zeeland en Friesland in de periode 2008-2012. In de figuren O.3 en O.4 zie je nog meer informatie over de verkochte eengezinswoningen in Zeeland en Friesland.

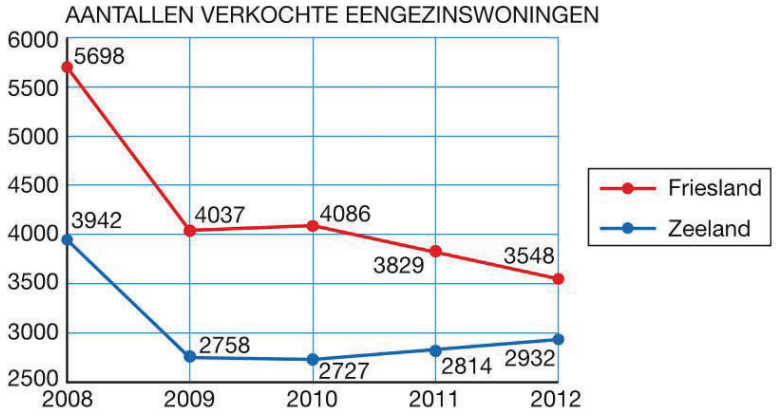


figuur O.2



figuur O.3

De totale marktwaarde van de in 2008 verkochte tussenwoningen in Zeeland bedroeg € 245 116 732,-. Door het inzakken van de woningmarkt is van deze woningen de gemiddelde waarde per woning in de periode 2008-2012 gedaald. Dat geldt ook voor de tussenwoningen die in 2008 in Friesland zijn verkocht, waarvan de totale waarde toen € 273 331 520,- bedroeg.

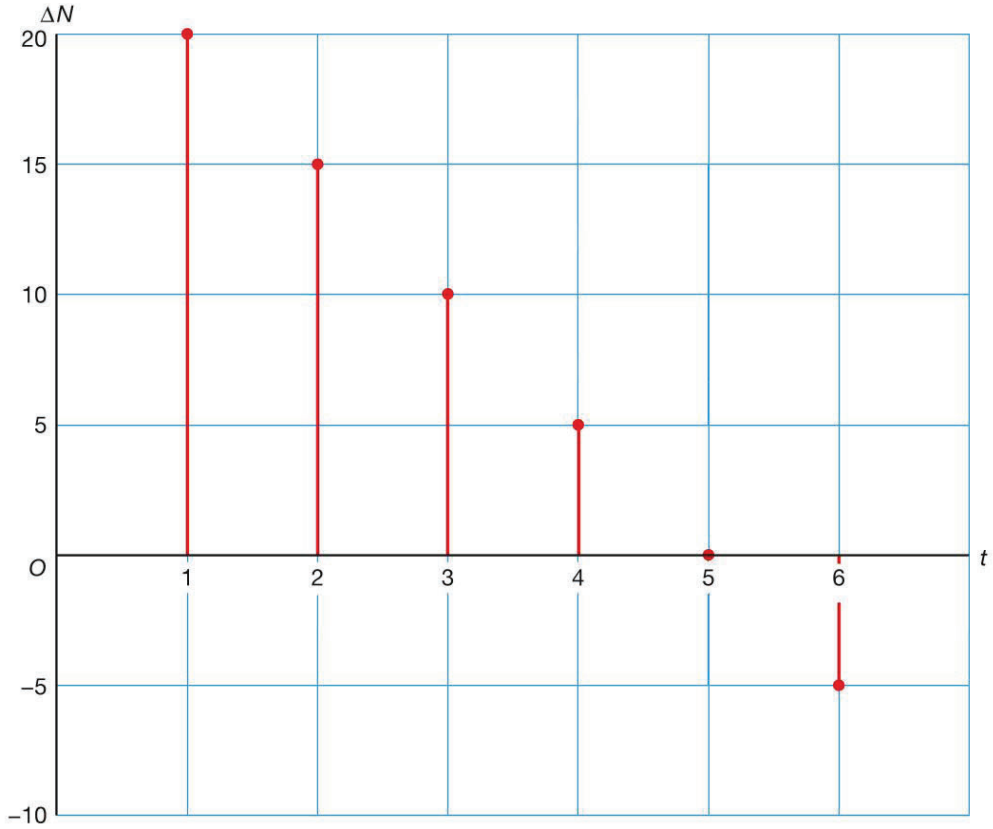


figuur O.4

Volgens Lois is de waarde-vermindering van de tussenwoningen in Friesland gemiddeld € 4500,- per woning meer dan in Zeeland. Onderzoek met behulp van de gegeven diagrammen of Lois gelijk heeft.

Voorspelling voorraad

De voorraadmanager van een scheepswerf houdt van alle onderdelen bij hoeveel er elke maand verbruikt worden. Van onderdeel A heeft hij de gegevens van het eerste halfjaar van 2014 verwerkt in het toenamediagram in figuur O.5. Hierin is N het aantal gebruikte onderdelen in een maand en t de tijd in maanden met $t = 0$ op 1 januari 2014. Ook is bekend dat er in maart 2014 van onderdeel A 1250 stuks zijn gebruikt.



figuur O.5

Directeur Feteris vraagt de voorraadmanager om een voorspelling te doen van het totale verbruik van onderdeel A in de tweede helft van 2014. De voorraadmanager overweegt drie mogelijke manieren.

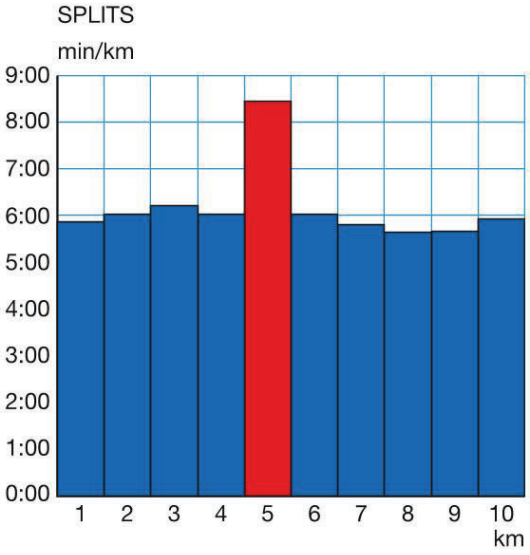
1 In de tweede helft van 2014 is het maandverbruik van onderdeel A iedere maand gelijk aan het maandverbruik in juni 2014.
2 In de tweede helft van 2014 is het totale verbruik van onderdeel A gelijk aan het totale verbruik in de eerste helft van 2014.
3 Het maandverbruik in de tweede helft van 2014 van onderdeel A neemt elke maand verder af volgens de trend in de eerste helft van 2014.

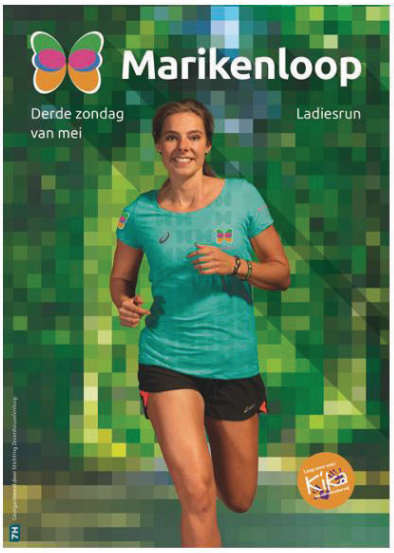
Bereken voor alle drie manieren het totale voorspelde verbruik van onderdeel A in de tweede helft van 2014 en bereken hoeveel procent de hoogste voorspelling meer is dan de laagste.

Marikenloop

De Marikenloop is een jaarlijkse hardloopwedstrijd voor vrouwen. Deze zogenaamde 'ladiesrun' is genoemd naar Mariken van Nimwegen, de bekendste Nijmeegse vrouw uit de geschiedenis. De wedstrijd gaat over vijf of tien kilometer en wordt op de derde zondag van mei gehouden in Nijmegen.
Marlijn heeft deelgenomen aan de 10 kilometer van de Marikenloop 2014. Haar snelheden per kilometer, ook wel 'splits' genoemd, zijn in de tabel en het bijbehorende diagram in figuur O.6 te zien.

km	tempo
1	5:52 min/km
2	6:02 min/km
3	6:12 min/km
4	6:01 min/km
5	8:27 min/km
6	6:02 min/km
7	5:48 min/km
8	5:39 min/km
9	5:40 min/km
10	5:55 min/km





figuur O.6

Bij een waterpost hebben lopers de mogelijkheid om water te drinken. Marlijn heeft tijdens haar 5e kilometer gebruik gemaakt van een waterpost en is tijdens het drinken gaan wandelen met een constante snelheid van 4 km/uur.
Na afloop wil Marlijn berekenen hoeveel meter ze heeft gewandeld vanaf de waterpost. Ze neemt daarbij aan dat ze de rest van de 5e kilometer met de gemiddelde snelheid van de overige 9 kilometer gelopen heeft.
Onderzoek of Marlijn volgens haar berekening meer of minder dan 250 meter heeft gewandeld.

Informatief Mariken van Nimwegen

Mariken van Nimwegen is een fictief personage uit een zogenaamd mirakelspel (een middeleeuwse vorm van toneel). Het mirakelspel over Mariken is een combinatie van waargebeurde en fictieve gebeurtenissen dat zich in de buurt van Nijmegen afspeelt.
Sinds 2009 wordt jaarlijks het mirakelspel Mariken van Nimwegen opgevoerd in de binnenstad van Nijmegen. Kunstenares Vera van Hasselt heeft van Mariken een beeld gemaakt dat op de grote markt in Nijmegen staat.

Hearthstone

Hearthstone is een digitaal verzamelkaartspel dat in maart 2014 is uitgebracht. Je begint het spel met een aantal gratis kaarten om mee te spelen. Daarna moet je pakjes kaarten kopen om je collectie uit te breiden. De kaarten zijn onderverdeeld in de vijf categorieën Free, Common, Rare, Epic en Legendary. De Legendary kaarten zijn het meest zeldzaam en daardoor erg gewild.
Het is voor spelers niet bekend wat de kans is dat er een Legendary kaart in een pakje kaarten zit. Deze kans is daardoor een veelbesproken onderwerp op de internetfora van het spel.
Op een van de internetfora wordt beweerd dat gemiddeld 1 op de 20 pakjes een Legendary kaart bevat. Fabio betwijfelt dit. Hij koopt regelmatig een pakje kaarten en heeft bijgehouden in hoeveel pakjes een Legendary kaart zat. Dat was in precies 10% van de pakjes zo. Hij stelt bij de proportie 10% het 95%-betrouwbaarheidsinterval op en ziet dat de relatieve frequentie $\frac{1}{20}$ buiten het interval ligt.
Voor zijn berekeningen maakt hij gebruik van de formule

$\sigma = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$, met σ de standaardafwijking, $\hat{p}$ de steekproefproportie en n het aantal pakjes kaarten in de steekproef.

Hoeveel pakjes kaarten heeft Fabio minstens gekocht om deze conclusie te kunnen trekken?

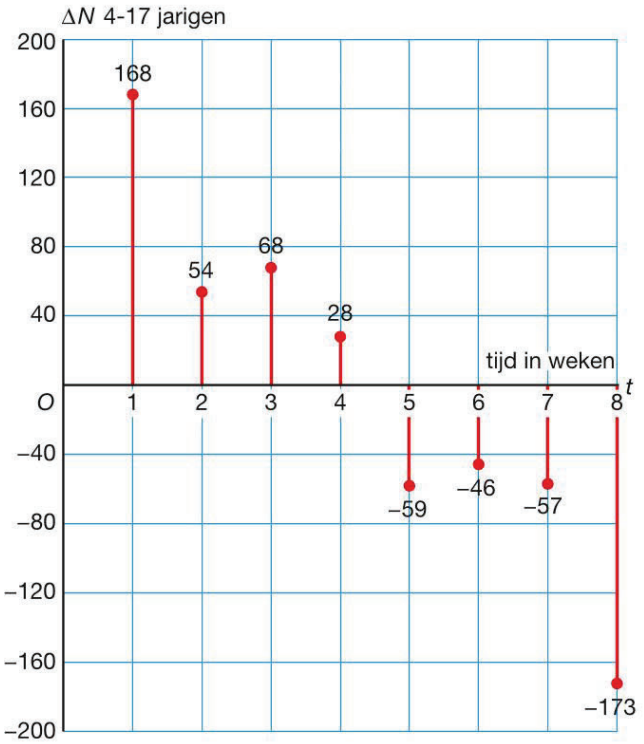
Camping in Zeeland

In Nederland hebben scholieren zes weken zomervakantie. De overheid heeft het land opgedeeld in de regio's Noord- , Midden- en Zuid-Nederland die elk een eigen vakantieperiode hebben.
De schoolvakantieperiode duurt daarom acht weken.

Deze schoolvakantieperiode is de drukste periode op camping Aan de Zeeuwse kust. De camping heeft een animatieteam dat activiteiten organiseert voor kinderen in de leeftijd van 4 tot en met 17 jaar. Iedere ochtend wordt er geknutseld en gedanst, 's middags en 's avonds zijn er veel leuke spelletjes en kinderen kunnen zich opgeven voor diverse sporttoernooien. Ook worden er disco's, speurtochten, musicals en modeshows georganiseerd.
Voor het plannen en organiseren van deze activiteiten houdt het animatieteam aan het begin van elke week bij hoeveel kinderen van 4 tot en met 17 jaar er op de camping zijn. Hiervan is een toenamediagram gemaakt. Zie de figuur hiernaast. Hierbij is t de tijd in weken met $t = 0$ aan het begin van de eerste vakantieweek.
Aan het begin van de tweede vakantieweek waren er 210 kinderen met een leeftijd van 4 tot en met 17 jaar.
Voor de kinderen van 0 tot en met 3 jaar is er geen activiteitenprogramma.
Gegeven is dat er aan het begin van de eerste vakantieweek 25 kinderen met een leeftijd van 0 tot en met 3 jaar op de camping waren. Na acht weken waren dat er 265. Ga uit van een lineaire toename van het aantal 0-3 jarigen in deze acht weken.
De camping hanteert voor het totaal aantal personen N_{totaal} op de camping het model

$$N_{\text{totaal}} = 3t^3 - 63t^2 + 405t + 120.$$

Bereken hoeveel kinderen er op de camping waren op het moment dat het totaal aantal personen maximaal was.
Waren er meer kinderen dan volwassenen? Hoeveel scheelt het?



figuur O.7

RUN Winschoten

Elk jaar wordt op de tweede zaterdag in september de RUN Winschoten gehouden. Dit hardloopfestijn bestaat uit verschillende onderdelen. Je hebt de zogenaamde Lutje RUN voor de jeugd, bijvoorbeeld 2 km voor de 13-, 14- en 15-jarigen. Ook heb je de RUN 50, dat is 50 km individueel, de RUN 100, dat is 100 km individueel en de ploegen 10 × 10 km estafette.
De RUN 100 is voor de echte ultralopers. Hieraan is het Nederlands kampioenschap 100 km verbonden, maar er doen ook veel lopers uit andere landen mee. Er wordt gelopen in tien rondes van elk 10 km.
In 2014 werd de 100 km gewonnen door de Tsjech Daniël Oralek in een tijd van 7 uur, 22 minuten en 35 seconden. Dit wordt genoteerd als 07:22:35.
De Nederlander Bram van Rijswijk eindigde als tweede en werd daarmee Nederlands kampioen op de 100 km.
In de tabel hieronder zie je dat de eindtijd van Bram 07:27:29 was. Bovendien zie je zijn tussentijden telkens na 10 km.

lap 1	lap 2	lap 3	lap 4	lap 5	lap 6	lap 7	lap 8	lap 9	lap 10	time
43:33	43:27	43:19	43:17	42:40	42:52	44:06	46:30	48:43	49:05	07:27:29

Uit de tabel volgt dat Bram over de eerste 70 km in elke ronde gemiddeld sneller liep dan zijn gemiddelde snelheid over de gehele afstand van 100 km.

Bereken hoeveel seconden voorsprong hij na 70 km had op de tijd die hij zou hebben gelopen als hij over de gehele afstand steeds met dezelfde gemiddelde snelheid zou hebben gelopen.