



STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Exploratieve statistiek

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Inhoudstafel

Een statistisch onderzoek naar de kleuren van M&M-snoepjes 1

1	Wat wil je weten? Hoe ga je meten?	1
1.1	De onderzoeksvraag	1
1.2	Een dataset maken	3
1.3	De dataset: getallen en context	5
2	Op speurtocht in de dataset	5
2.1	Een frequentietabel opstellen	5
2.2	Figuren tekenen	7
2.2.1	Een staafdiagram	7
2.2.2	Een taartdiagram	9
3	Wat heb je gevonden? Hoever kan je gaan in je conclusie?	11
3.1	De variabiliteit van steekproefresultaten	11
3.2	Steekproefgrootte, nauwkeurigheid en haalbaarheid	12
3.3	Een model voor de populatie	13
4	Kernachtige samenvatting van dit onderzoek	15
5	Zelfevaluatie	16

Een statistisch onderzoek naar de mening van leerlingen over het uur van thuiskomst na een avondje uit 19

1	Wat wil je weten? Hoe ga je meten?	19
1.1	De onderzoeksvraag	19
1.2	Een dataset maken	21
1.3	De dataset: getallen en context	21
2	Op speurtocht in de dataset	23
2.1	De frequentietabel	23
2.2	Het staafdiagram	24
3	Wat heb je gevonden? Hoever kan je gaan in je conclusie?	25
3.1	De variabiliteit van steekproefresultaten	25
3.2	Een enquête afnemen	26
3.3	Wat kan er mis gaan?	27
3.3.1	Problemen met de vraag	27
3.3.2	Problemen met de respons	28
3.3.3	Problemen met de selectie van de deelnemers	28
4	Kernachtige samenvatting van dit onderzoek	30
5	Zelfevaluatie	31

Een statistisch onderzoek naar honden en katten in een gezin..... 34

1	Wat wil je weten? Hoe ga je meten?	35
1.1	De onderzoeksvraag	35
1.2	De dataset: getallen en context	36
2	Op speurtocht in de dataset.....	36
2.1	Een frequentietabel opstellen.....	37
2.2	Een staafdiagram tekenen	38
2.3	Numerieke kenmerken: gemiddelde en mediaan	41
2.4	Een staafdiagram interpreteren	42
3	Wat heb je gevonden? Hoever kan je gaan in je conclusie?.....	43
3.1	De variabiliteit van steekproefresultaten	43
3.2	Een uitspraak over de populatie	43
4	Kernachtige samenvatting van dit onderzoek	44
5	Zelfevaluatie	45

Een statistisch onderzoek naar het schatten van de tijdsduur van 1 minuut..... 51

1	Wat wil je weten? Hoe ga je meten?	51
1.1	De onderzoeksvraag	51
1.2	De dataset: getallen en context.	52
2	Op speurtocht in je dataset.....	52
2.1	Een frequentietabel met klassenindeling	53
2.2	Het histogram	55
2.3	Numerieke kenmerken	56
2.3.1	Gemiddelde en mediaan	56
2.3.2	Standaardafwijking en interkwartielafstand	56
2.4	De boxplot	57
2.5	Histogram en boxplot interpreteren.....	58
3	Wat heb je gevonden? Hoever kan je gaan in je conclusie?.....	60
3.1	De variabiliteit van steekproefresultaten	60
3.2	Enkelvoudig aselekt, en nog veel meer	60
3.3	Een uitspraak over de populatie	61
4	Kernachtige samenvatting van dit onderzoek	61
5	Zelfevaluatie	62

Een statistisch onderzoek naar de kleuren van M&M-snoepjes

Je eerste onderzoek

Hier komt je allereerste statistisch onderzoek. Spannend! Maar hoe zit zo'n statistisch onderzoek eigenlijk in elkaar?

De manier van werken kan je in vier stappen samenvatten:

1. Wat wil je weten? Hoe ga je meten?
2. Op speurtocht in de dataset.
3. Wat heb je gevonden? Hoe ver kan je gaan in je conclusies?
4. Kernachtige samenvatting van je onderzoek.

Wat die vier stappen precies inhouden, dat leer je door zelf enkele onderzoeken uit te voeren.

1 Wat wil je weten? Hoe ga je meten?

1.1 De onderzoeksvraag

Iedereen kent wel M&M's, de chocoladesnoepjes met de felgekleurde suikerjasjes. De fabrikant van M&M's stopt verschillende kleuren snoepjes in één verpakking.

Heb je enig idee welke kleuren allemaal voorkomen bij M&M's? Komt elke kleur evenveel voor? Dat ga je nu onderzoeken.

Je hebt hier al een eerste probleem. Wat wil je eigenlijk onderzoeken? Wil je iets zeggen over de kleuren in je eigen zakje M&M's of wil je iets zeggen over de kleuren van alle M&M-snoepjes die door de fabrikant gemaakt worden? Dat zijn nogal verschillende vragen!



Om goed het onderscheid te maken tussen “alle M&M's” en “de snoepjes in jouw zakje M&M's” gebruikt de statistiek twee verschillende woorden. Je spreekt over **populatie** als je “de totale verzameling” bedoelt (dus alle M&M-snoepjes). Meestal heb je geen tijd of geld om een volledige populatie te onderzoeken en daarom bekijk je enkel een klein deeltje van die populatie. Zo'n deeltje van een populatie wordt in de statistiek een **steekproef** genoemd. De snoepjes die in je zakje M&M's zitten, zijn een heel klein deeltje van alle M&M's. Jouw snoepjes zijn dus een steekproef uit de totale populatie van alle M&M's.

Je steekproef bestaat uit dingen die je zelf hebt verzameld, die je dus zelf kan zien en beschrijven (met getallen en grafieken). Hoe je dat doet, dat ga je in dit onderzoek leren. Maar misschien wil je daarna ook iets zeggen over alle M&M's. Misschien zijn de blauwe snoepjes in jouw zakje in de meerderheid. Zou je dan kunnen zeggen dat bij alle M&M's de blauwe snoepjes het meest voorkomen? (Let op! Misschien heeft een andere leerling meer rode snoepjes).

Iets zeggen over de totale populatie als je enkel de steekproef ziet, dat is helemaal niet eenvoudig. Statistiek kan je hierbij helpen. Een eerste hulp die de statistiek je biedt, gaat over de manier waarop je een steekproef moet trekken. De raadgeving die je hier krijgt, had je waarschijnlijk nooit verwacht. Om een goede steekproef te trekken, moet je je laten leiden door ... het toeval!

Je laten leiden door het toeval, dat is gemakkelijker gezegd dan gedaan. Dat zal je ondervinden in je volgende onderzoeken. Maar vandaag gaat het over M&M's. Die worden gemaakt in verschillende kleuren volgens een verhouding die door de fabrikant is vastgelegd. Die snoepjes komen terecht in een reuzegrote container waar ze grondig door elkaar worden gemengd. Daarna wordt uit die container lukraak een schep snoepjes genomen en die snoepjes worden in een zakje verpakt. Dat gebeurt natuurlijk allemaal volautomatisch en in superhygiënische omstandigheden.

Die enorme container, waarin miljoenen M&M's zitten, kan je beschouwen als een goed model voor de hele populatie. Een goede steekproef trek je dan als volgt: "goed mengen en dan lukraak trekken". Deze manier van werken krijgt in de statistiek de naam "**enkelvoudige aselechte steekproef**". Het aantal elementen in je steekproef (het aantal getrokken snoepjes) noteer je door de letter " n " (dat noem je de **steekproefgrootte**).

Als je iets over de kleuren van de hele populatie van M&M's wil weten, dan kan je ook als volgt te werk gaan. Trek lukraak een snoepje uit de goed gemengde container. Noteer de kleur van het getrokken snoepje en leg het dan terug in de container. Meng terug goed en herhaal dit nu 50 keer. Op die manier heb je ook 50 keer een kleur genoteerd. Als je zo werkt, dan spreek je over "**trekken met terugleggen**". Als je alle snoepjes bijhoudt, dan spreek je over "**trekken zonder terugleggen**". Eigenlijk maakt het niet zoveel verschil of er nu 50 snoepjes meer of minder zitten in een goed gemengde container met miljoenen snoepjes.

De meeste steekproeven die je in de praktijk tegenkomt zijn van het type "trekken zonder terugleggen". Zolang je steekproef veel kleiner is dan de totale populatie hoef je hier geen extra aandacht aan te besteden. Als vuistregel zorg je er voor dat je steekproef niet groter is dan 10% van de totale populatie.

- *Snoepjes kan je echt in een grote container gooien en door elkaar mengen. Maar hoe zou jij een enkelvoudige aselechte steekproef trekken uit de populatie van alle leerlingen van je school?*

1.2 Een dataset maken

De informatie in je steekproef ga je nu op een overzichtelijke manier opschrijven. Zo krijg je **de gegevensverzameling of dataset**.



Lees in je infoboekje “De structuur van een dataset” voor je verder werkt.

Denk nu goed na hoe jij dit onderzoek over de kleuren van M&M's gaat uitvoeren.

- *Welke gegevens ga je noteren?*
- *Maak een tabel waarin je de gegevens zal opschrijven voor jouw zakje M&M's. Begin met een kleine tabel voor een viertal snoepjes en overleg met je leerkracht of de tabel die je zo opstelt goed is. Als je afkortingen gebruikt, schrijf dan ook op wat die afkortingen betekenen.*

- *Maak nu een grote tabel om alle gegevens te noteren die je opmeet bij het onderzoek van jouw zakje M&M's. Noteer ook duidelijk de titel, de datum en je naam. De dataset die je zo opstelt, vormt de basis voor al je verder onderzoek.*

- *De snoepjes in je zakje kan je bekijken als een steekproef uit alle M&M's. Is deze steekproef getrokken met terugleggen of zonder terugleggen?*
- *Wat is jouw steekproefgrootte en hoe noteer je die?*
- *Wat zijn voor uw dataset de elementen?*
- *Welke veranderlijke heb je bij die elementen genoteerd?*

1.3 De dataset: getallen en context

Bij de dataset die je pas hebt opgesteld is er één kolom waarin je de kleur van de snoepjes hebt geschreven. Je hebt hier te maken met een “eigenschap van snoepjes”, namelijk “hun kleur”. Dit is een “veranderlijke” die jij hebt opgemeten. Deze veranderlijke heeft hier de waarden: rood, groen, blauw, bruin, en geel.

Op kleuren kan je geen zinvolle wiskundige bewerkingen uitvoeren zoals optellen of vermenigvuldigen. Daarom noemt men de veranderlijke “kleur” een **kwalitatieve** veranderlijke.

Als je kleuren hebt, zoals rood en groen, dan kan je even goed eerst groen zeggen en dan rood, in plaats van eerst rood en dan groen. Er is geen enkele reden waarom de ene volgorde beter is dan de andere. Enkel de naam van de kleur is van belang en daarom noemt men zo’n veranderlijke **nominaal**. De “kleur” van een snoepje is dus een **nominaal kwalitatieve veranderlijke**.



Er is een belangrijk onderscheid tussen de naam van een veranderlijke en de verschillende waarden van die veranderlijke.

In dit geval is “kleur” de **naam** van de veranderlijke en “rood, groen, blauw, ...” zijn de mogelijke **waarden**.

2 Op speurtocht in de dataset

Je dataset is de basis voor al je verder onderzoek. De dataset, samen met de beschrijving van hoe je hem hebt opgemeten, moet je nauwkeurig bewaren.

2.1 Een frequentietabel opstellen

- Gebruik je dataset om een frequentietabel op te stellen. Doe dat zoals hieronder aangegeven.

In de eerste kolom schrijf je de kleuren en in de tweede kolom schrijf je hoeveel snoepjes er van die kleur zijn. Dit aantal heet de **frequentie** van die kleur. Een tabel die je op deze manier opstelt, heet een **frequentietabel**. Zorg ervoor dat je elke kolom een juiste naam geeft: deze naam schrijf je bovenaan de kolom.

- Hoe kan je de steekproefgrootte snel berekenen met behulp van de frequentietabel?

We gaan nu een derde kolom aan de frequentietabel toevoegen. In die kolom komt, per kleur, **de relatieve frequentie**. De relatieve frequentie is niets anders dan de frequentie gedeeld door het totale aantal n . Je kan dit getal ook in percent uitdrukken. Als je voor “geel” een relatieve frequentie van 0.16 vindt, dan kan je dat ook schrijven als 16%. Hierbij rond je af op één eenheid. In woorden zeg je dat 16 % van jouw onderzochte snoepjes geel is.

Om zoveel mogelijk van je tijd te kunnen besteden aan nadenken en discussiëren, ga je zo weinig mogelijk tijd besteden aan slaafse berekeningen. Gebruik je GRM op een verstandige manier. Zo leer je ook hoe elk “echt” statistisch onderzoek verloopt.

Als je GRM lijsten bevat die je nog nodig hebt, bewaar die dan eerst. Start met voldoende vrij geheugen. Herstel de standaardlijsten: druk **[STAT]**, kies 5:SetUpEditor en dan **[ENTER]**.

Zet nu de frequenties in lijst [L1]. Als je bijvoorbeeld 13 rode, 9 groene, 8 gele, 7 oranje, 7 bruine en 6 blauwe snoepjes had, dan ga je als volgt te werk. Druk **[STAT]** en kies 1:Edit... . Je komt dan in de lijsten terecht. Daar kan je de frequenties gewoon in lijst [L1] intikken. Na elk getal druk je **[ENTER]**. Kijk of je alles goed hebt ingetikt. Nu ga je alle frequenties in [L1] delen door het totaal aantal getallen en het resultaat in [L2] plaatsen. Zo krijg je de relatieve frequenties in [L2].

Je kan zelf tellen hoeveel snoepjes je hebt (bijvoorbeeld 50) en dan zeggen dat alle getallen in [L1] door 50 moeten gedeeld worden. Maar je weet dat de som van alle frequenties gelijk is aan het totaal aantal. Dus kan je ook zeggen dat de getallen in [L1] moeten gedeeld worden door “de som van alle frequenties”. Die frequenties staan in [L1] en dus deel je door de som van de getallen in [L1]. Doe dit nu als volgt. Ga op de kop van [L2] staan en druk **[ENTER]**. Vervolledig het commando $[L2] =$ met **[2nd] [L1] [÷] [2nd] [LIST]** en loop met het pijltje **[→]** naar MATH en kies dan 5:sum(. Druk dan **[2nd] [L1] [)]** en **[ENTER]**. Kijk wat er in [L2] staat. In dit voorbeeld zijn de relatieve frequenties 26%, 18%, 16%, 14%, 14% en 12%.

```

3001 CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

```

```

SetUpEditor      Done

```

L1	L2	L3	1
13	-----	-----	
9			
8			
7			
7			
6			

L1(?)=

L1	L2	L3	2
13	-----	-----	
9			
8			
7			
7			
6			

L2=L1/sum(L1)

L1	L2	L3	3
13	.26	-----	
9	.18		
8	.16		
7	.14		
7	.14		
6	.12		

L3(?)=

- Voeg aan je tabel een derde kolom toe met naam “relatieve frequentie” en schrijf daarin de resultaten die in [L2] staan (in percent). Tel de percenten bij elkaar op. Hoeveel heb je?

Soms heb je frequenties nodig, in andere gevallen gebruik je relatieve frequenties. Als je aantallen bestudeert, dan werk je met frequenties. Als je percentages gebruikt om twee onderzoeken met elkaar te vergelijken, dan werk je met relatieve frequenties.

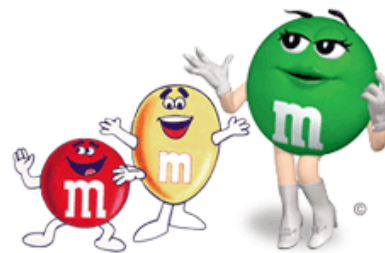
Voorbeeld

Om te weten of je genoeg rode snoepjes hebt om er eentje te kunnen geven aan elk van je 10 vrienden, dan kijk je naar de frequentie.

Als je de kleurensamenstelling van een grote en een kleine zak M&M's wil vergelijken, dan zal je met percentages werken en dus relatieve frequenties gebruiken.

2.2 Figuren tekenen

Veruit het meest belangrijke onderdeel bij de studie van een dataset is kijken naar figuren. Dit is niet eenvoudig en je moet stapsgewijs leren waar je allemaal moet op letten. Zodra je dit wat kent, kan je uit een figuur heel veel informatie halen. Maar je moet natuurlijk eerst weten welke figuur je moet maken en hoe je die moet tekenen.



2.2.1 Een staafdiagram

Je hebt in dit onderzoek een nominaal kwalitatieve veranderlijke opgemeten. Voor dit soort veranderlijken is het **staafdiagram** de basisfiguur.

Als voorbeeld zie je hier een staafdiagram van de Vlaamse bevolking per provincie. De informatie die hier wordt weergegeven, kan je vinden in het boekje “Vlaanderen in cijfers” op de website:

<http://aps.vlaanderen.be/statistiek/publicaties/pdf/vic/vic2005.pdf> . De namen van de provincies zijn afgekort als:

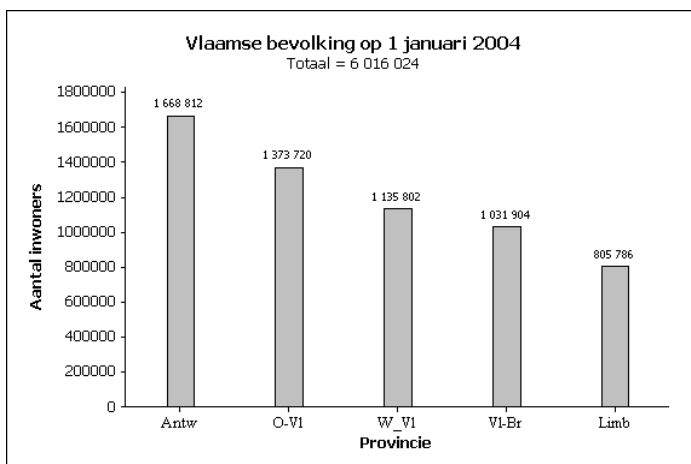
Antw = Antwerpen,

O-Vl = Oost-Vlaanderen,

W-Vl = West-Vlaanderen,

Vl-Br = Vlaams - Brabant,

Limb = Limburg.



Om een staafdiagram te tekenen op basis van jouw frequentietabel begin je als volgt.

- *Op de x-as zet je de verschillende kleuren. Hoewel de waarden van een nominale veranderlijke geen natuurlijke volgorde hebben, zal je toch moeten kiezen hoe je de kleur ordent op de x-as. Welke kleur komt als eerste? Welke kleur komt als tweede? Waarom maak je die keuze?*

- *Op de y-as duid je de frequentie van elke kleur aan en je tekent dan bij elke kleur een staafje waarvan de lengte overeenkomt met de frequentie van die kleur. Zorg ervoor dat alle staafjes los van elkaar staan.*
- *Voorzie de assen van de juiste naam.*
- *Teken nu zo'n staafdiagram voor jouw onderzoek.*



2.2.2 Een taartdiagram

Misschien wil je de *relatieve* frequenties van de kleuren grafisch voorstellen. Dan kan je ook een staafdiagram tekenen, waarbij je in de y-richting staafjes tekent waarvan de lengte gelijk is aan de relatieve frequentie.

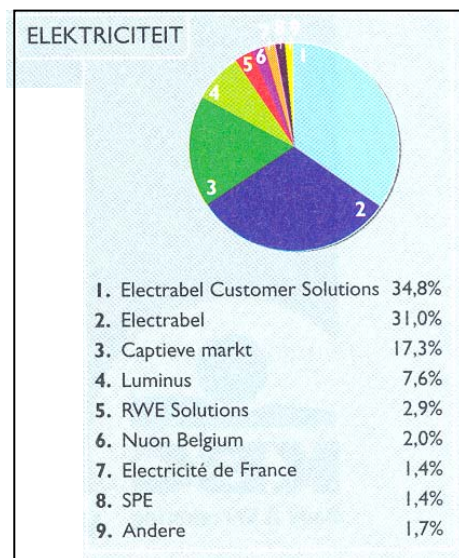
Maar er is ook een andere figuur die de relatieve frequentie (meestal uitgedrukt in percent) mooi weergeeft. Dat is het **taartdiagram** of **cirkeldiagram**.

Bij het tekenen van een taartdiagram verdeel je een cirkeloppervlak in stukken, juist zoals je een taart in stukken snijdt. Zo'n stuk heet een **sector**. De totale oppervlakte van de cirkel komt overeen met de som van alle percentages en dat is 100 %.

Voor een taartdiagram maken we enkele afspraken:

- Begin bovenaan en draai naar rechts
- De grootste sector komt eerst, dan komt de tweede grootste, enzovoort.

Je ziet hier een voorbeeld van de marktaandelen van energiebevoorraders in België. Het gaat over de elektriciteit in het jaar 2004. Deze figuur staat in het weekblad Knack van 22 juni 2005 en is goed leesbaar. Maar als je een krant of weekblad doorbladert, dan zie je soms verwarrende en zelfs verkeerde grafieken.



Hoeveel graden elke sector is, bereken je door de relatieve frequentie te vermenigvuldigen met 360°. Dit doe je met je GRM. Maak daarna “verstandige” afrondingen zodat alle sectoren samen terug 360° geven (je kan eventueel enkele keren tot op een halve graad werken).

L2*360÷L3			
L1	L2	L3	1
13	.26	93.6	
9	.18	64.8	
8	.16	57.6	
7	.14	50.4	
7	.14	50.4	
6	.12	43.2	
L1(7)=			

Druk **[2nd] [L2] [×] 360 [STO] [2nd] [L3] [ENTER]**.

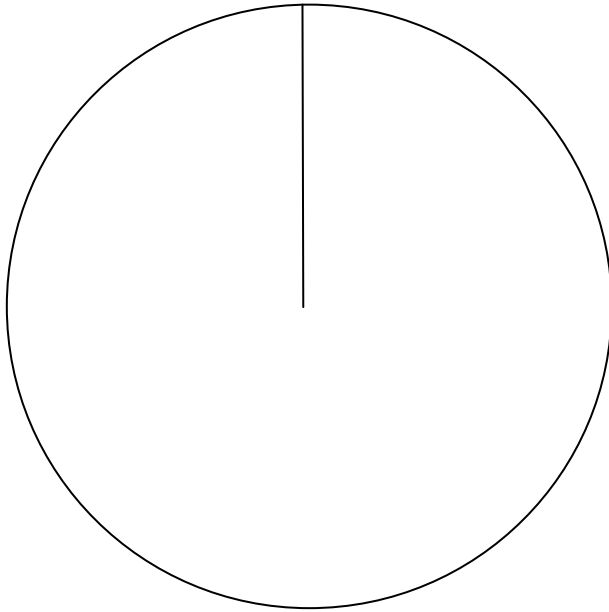
Bekijk de lijst **[L3]** met **[STAT]** en kies 1:Edit... . In dit voorbeeld heb je opeenvolgende sectoren van 94°, 65°, 58°, 50°, 50° en 43°.

Om een taartdiagram te tekenen op basis van jouw onderzoek begin je als volgt.

- *Bereken eerst voor elke relatieve frequentie hoe groot de sector is die daarbij hoort. Gebruik je GRM.*

KLEUR						
HOEK						

- *Teken nu cirkelsectoren die overeenstemmen met de relatieve frequenties. Schrijf bij elke sector met welke kleur van snoepje hij overeenkomt en noteer ook de relatieve frequentie in percentvorm erbij. Je kan natuurlijk ook de sector inkleuren met de bijhorende kleur.*



3 Wat heb je gevonden? Hoever kan je gaan in je conclusie?

3.1 De variabiliteit van steekproefresultaten

Je hebt nu de kleur van de snoepjes in jouw zakje M&M's bestudeerd met behulp van een frequentietabel, een staafdiagram en een taartdiagram. Je medeleerlingen hebben ook zo'n onderzoek gedaan voor de snoepjes die zij hebben gekregen.

- *Verwacht je dat je medeleerlingen dezelfde resultaten hebben gevonden als jij?*
- *Kan je je antwoord op vorige vraag wat verduidelijken door te verwijzen naar de manier waarop die zakjes gevuld worden? Kan je hierbij ook de woorden populatie en steekproef op een juiste wijze gebruiken?*

In plaats van naar alle kleuren te kijken, zou je er eens je lievelingskleur kunnen uithalen, bijvoorbeeld blauw. Hoeveel percent blauwe snoepjes zaten er in jouw zakje? En hoeveel percent blauwe snoepjes waren er bij je klasgenoten?

- *Noteer voor elk onderzocht zakje in je klas telkens het percent blauwe snoepjes.*
- *Als jij alleen maar je eigen zakje snoepjes mag onderzoeken en je zou moeten raden hoeveel percent blauwe snoepjes er door de fabrikant gemaakt wordt (dus hoeveel percent blauwe snoepjes er in de totale populatie zit), wat zou jij dan antwoorden?*
- *Is je bovenstaand antwoord exact juist? Hoe weet je dat?*

3.2 Steekproefgrootte, nauwkeurigheid en haalbaarheid

Met het onderzoek van de snoepjes in een aantal zakjes M&M's wil je een zicht krijgen op alle M&M-snoepjes. Je zou bijvoorbeeld willen weten hoeveel percent van alle M&M's blauw zijn of op welke manier de kleuren verdeeld zijn.

Een eerste (maar naïeve) reactie zou kunnen zijn: wel, onderzoek dan de totale populatie. Maar dat voorstel is helemaal niet haalbaar! Je gaat toch niet alle zakjes openmaken om te kijken wat de kleur van de snoepjes is. Dat zou niet alleen veel te veel tijd en geld vragen, het is gewoon onmogelijk omdat de snoepjes dan niet meer kunnen verkocht worden. Daarom onderzoek je dus maar een beperkt aantal snoepjes: je verzamelt informatie over een deel van de M&M's om zo conclusies te trekken over alle M&M's.

Herinner je de twee belangrijke begrippen:

- De hele groep objecten (of personen) waarover je iets wil weten, heet de **populatie**.
- Een **steekproef** is een deel uit deze populatie.

Als het praktisch haalbaar is en als je op een goede manier steekproeven trekt, dan is het beter om met een grotere steekproef te werken dan met een kleinere. Intuïtief kan je dit waarschijnlijk wel begrijpen. Als je een groter aantal M&M's uit de totale populatie mag trekken, dan heb je meer informatie. Maar ook een grote steekproef is nog altijd aan het toeval onderhevig. Als je echter meerdere keren een grote steekproef zou trekken, dan zou je zien dat op grotere steekproeven minder schommelingen zitten dan op kleinere.

Om een grotere steekproef te krijgen, kan je alle M&M's uit je klas samenbrengen in één grote steekproef.

- *De verschillende resultaten van elk onderzocht zakje worden nu verzameld. Noteer alle cijfers die op bord komen en maak dan een nieuwe frequentietabel met daarin per kleur de frequentie en de relatieve frequentie voor de snoepjes van de totale klas.*

<i>Kleur</i>	<i>Frequentie</i>	<i>Relatieve frequentie</i>

3.3 Een model voor de populatie

Hoe de echte populatie van alle M&M-snoepjes eruitziet, zal niemand ooit weten. Je kan toch niet naast die productielijn gaan staan en voor die miljoenen (miljarden?) snoepjes de kleur noteren. Maar een **model** voor de populatie bestaat wel. In België worden geen M&M's gemaakt. Zij worden ingevoerd uit naburige landen. Bij M&M's uit Frankrijk komen alle kleuren globaal (in de populatie) evenveel voor. Dat staat op hun website: <http://www.m-ms.fr/front/fr-fr/html/index.html>.



Als je nu mag aannemen dat (volgens de fabrikant) alle kleuren evenveel voorkomen dan kan je deze eigenschap gebruiken om een model te maken voor de totale populatie. Aan de andere kant heb jij nu cijfers van een grote steekproef uit die populatie. Het is dus interessant om het model voor de populatie te vergelijken met wat je ziet in die grote steekproef. Je zou hiervoor twee afzonderlijke staafdiagrammen kunnen tekenen maar je kan die vergelijking ook in één en dezelfde figuur voorstellen.



Als je de resultaten van meerdere situaties in eenzelfde staafdiagram toont, dan spreek je over een **staafdiagram met subgroepen**.

Bij zo'n staafdiagram hoort altijd een legende, waarin je aangeeft welke kleur of arcering bij welke subgroep hoort.

- Maak een tabel waarin je aangeeft hoe de populatie er precies uitziet. Gebruik je frequenties of relatieve frequenties in die tabel?

Kleur	

- *Hoe ga je de grafiek met subgroepen tekenen? Welke volgorde kies je voor de kleuren op de x-as?*

- *Teken nu de grafiek.*



- *Kan je verklaren waarom jouw cijfers eventueel afwijken van die van de fabrikant?*

4 Kernachtige samenvatting van dit onderzoek

Een statistisch onderzoek wordt niet zomaar in het wilde weg gedaan. Meestal is er een opdrachtgever (bedrijf, overheid, organisatie, ...) die bepaalde informatie nodig heeft. De statisticus die het onderzoek heeft uitgevoerd zal dan ook zijn onderzoeksresultaten zorgvuldig moeten presenteren bij die opdrachtgever.

Op dit ogenblik heb je al heel wat informatie over het onderzoek. Deze informatie moet je nu nog vervolledigen met:

- *antwoorden op de contextvragen*
- *besluiten over het uitgevoerde onderzoek.*

De contextvragen of www-vragen, die bij elk onderzoek aan bod komen, zijn:

1. **Waarom** is dit onderzoek uitgevoerd? (*Wie wilt wat weten?*)
2. **Waar** is dit onderzoek uitgevoerd? (*In het buitenland? In mijn gemeente?*)
3. **Wanneer** is dit onderzoek uitgevoerd? (*Vorige eeuw? Dit jaar?*)
4. **Wie** wordt onderzocht? (*Bij wie worden dingen opgemeten? Wat zijn de “elementen”?*)
5. **Wat** wordt er juist opgemeten? (*Wat wordt er per element allemaal genoteerd? Wat zijn de “veranderlijken”?*)
6. **Hoe** wordt dit onderzoek uitgevoerd? (*Hoe zijn de “elementen” verzameld? Hoe zijn mensen bij een enquête gecontacteerd?*)

- *Formuleer in een bondige tekst de antwoorden op de contextvragen.*

Nu kan je conclusies trekken. Maar je hebt al begrepen dat statistische besluiten rekening moeten houden met toevallige uitkomsten en dus niet hetzelfde zijn als wiskundige bewijzen.

Wees dus voorzichtig bij je besluit. Als er problemen zijn opgetreden, vermeld die dan. Zo kom je tot een genuanceerd rapport.

- *Formuleer in een bondige tekst je besluiten over het uitgevoerde onderzoek.*



5 Zelfevaluatie

In dit onderzoek heb je geleerd over:

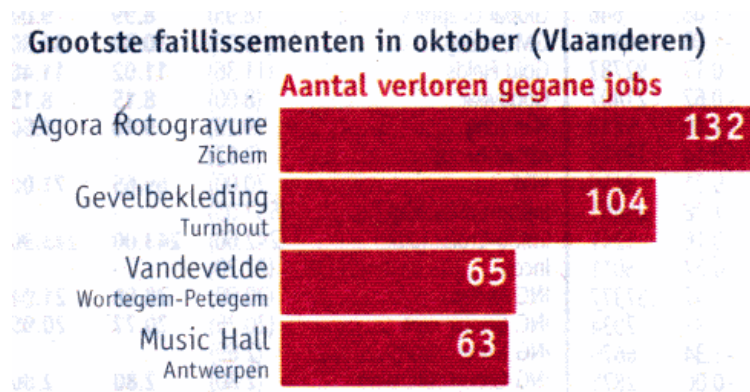
- de context van een statistisch onderzoek (wanneer, waar,...)
- het onderscheid tussen de populatie en een steekproef
- een enkelvoudige aselechte steekproef
- de structuur van een dataset (elementen, veranderlijken)
- nominaal kwalitatieve veranderlijken
- de frequentietabel bij een nominaal kwalitatieve veranderlijke
- het staafdiagram bij een nominaal kwalitatieve veranderlijke
- het taartdiagram bij een nominaal kwalitatieve veranderlijke
- het staafdiagram met subgroepen.



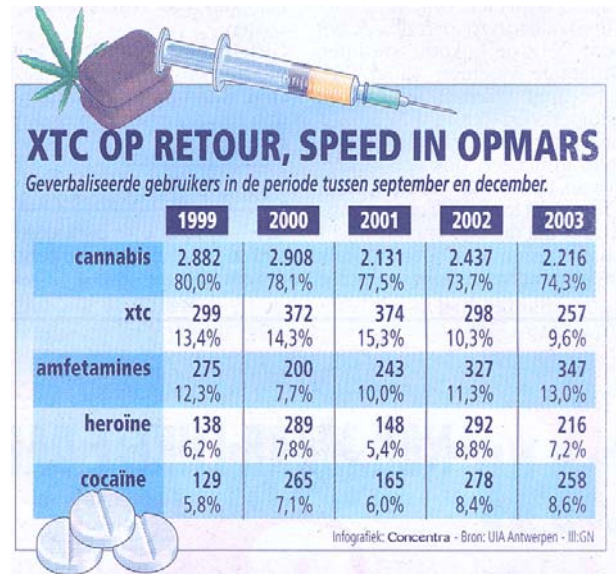
Je bent nu in staat om de volgende opdrachten uit te voeren:

- *Zeg in eigen woorden op welke vragen je een antwoord moet kunnen geven als men vraagt naar de context van een statistisch onderzoek.*
- *Omschrijf de begrippen steekproef en populatie in je eigen woorden en geef een (nieuw) voorbeeld. Leg voor jouw voorbeeld uit hoe je daar een enkelvoudige aselechte steekproef zou trekken.*
- *Leg duidelijk uit hoe een dataset eruitziet. Gebruik hiervoor een nieuw voorbeeld dat je zelf hebt bedacht.*

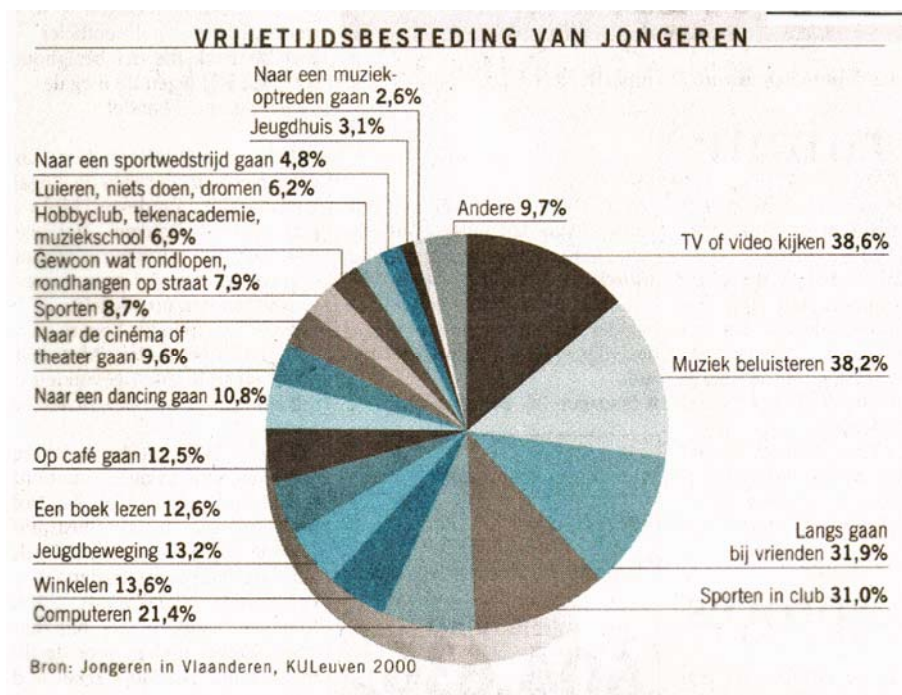
- Zeg in eigen woorden wanneer je van een kwalitatieve veranderlijke zegt dat ze nominaal is. Is de bloedgroep zo'n veranderlijke? Kan je zelf een nominaal kwalitatieve veranderlijke bedenken?
- Als je opmetingen hebt van een nominaal kwalitatieve veranderlijke, dan moet je daarvoor een frequentietabel, een staafdiagram en een taartdiagram kunnen maken.
- Soms kom je de uitdrukking “horizontaal staafdiagram” tegen. Kijk daarvoor naar de figuur die je vindt in de Gazet van Antwerpen van 3 november 2004. Voor de 364 jobs die in oktober 2004 bij de 4 grootste faillissementen in Vlaanderen verloren gingen, heeft men een figuur getekend. Welke veranderlijke is er genoteerd bij elke persoon die zijn job is kwijtgeraakt? Welk soort veranderlijke is dat? Wat zijn haar waarden? Is de figuur goed getekend? Leg nu in je eigen woorden uit wat een horizontaal staafdiagram is en wanneer je zoiets gebruikt.



- Als je een frequentietabel ziet, dan moet je die juist kunnen interpreteren. Bekijk de tabel over XTC en Speed (Gazet van Antwerpen van 20-10-2004). Kijk enkel naar de informatie die daarin staat over het jaar 2003. Is “druggebruik” daar behandeld als een nominaal kwalitatieve veranderlijke? Is de tabel correct?



- Een bestaande figuur moet je juist kunnen interpreteren. Bekijk het taartdiagram over de vrijetijdsbesteding van jongeren (De Standaard van 6-12-2000). Is “vrijetijdsbesteding” hier behandeld als een nominaal kwalitatieve veranderlijke? Is het taartdiagram correct getekend?



Een statistisch onderzoek naar de mening van leerlingen over het uur van thuiskomst na een avondje uit

1 Wat wil je weten? Hoe ga je meten?

1.1 De onderzoeksvraag

Een methode die heel vaak gebruikt wordt om informatie te verzamelen is de **enquête**. Een enquête is een onderzoeksmethode waarbij je een lijst van zorgvuldig geselecteerde vragen gebruikt. Je kan hierbij een onderscheid maken tussen **open** vragen en **gesloten** vragen.

Bij een open vraag mag je het antwoord formuleren in je eigen woorden. Voorbeelden hiervan zijn:

- wat moet er volgens jou veranderen aan het schoolreglement?
- wat versta je onder een gezonde voeding?
- hoeveel zou volgens jou een combiticket voor Pukkelpop mogen kosten?

Bij gesloten vragen mag je enkel kiezen uit vooraf vastgelegde mogelijkheden:

- hoeveel auto's hebben jullie thuis?
 - ☐ 0
 - ☐ 1
 - ☐ 2
 - ☐ meer dan 2

In een **opinieonderzoek** probeer je te weten te komen wat mensen denken over bepaalde onderwerpen. Je hebt misschien zelf al ondervonden dat de meeste dingen in het leven niet zomaar simpelweg te klasseren zijn als “goed of slecht”, “zwart of wit”, “ja of neen”. Dikwijls zijn er heel wat schakeringen tussenin. Daarom zal men in een opinieonderzoek vaak een bewering formuleren waarbij de **respondent** moet aangeven in hoeverre hij daarmee akkoord gaat. Een voorbeeld zou kunnen zijn:

- de huidige regering maakt haar beloften waar
 - ☐ helemaal akkoord
 - ☐ akkoord
 - ☐ niet akkoord
 - ☐ helemaal niet akkoord
 - ☐ geen mening

In dit tweede onderzoek ga je nu zelf een enquête gebruiken voor een opinieonderzoek. Je wil immers weten wat je leeftijdsgenoten denken over het uur van thuiskomst na een fuif of een avondje uit. Eigenlijk wil jij weten wat zij denken over “wie er mag beslissen over het uur van thuiskomst”: hun ouders of zichzelf.

Je zou je leeftijdsgenoten kunnen vragen om bij de onderstaande bewering één antwoord aan te kruisen:

Bewering: “Tot nu toe vind ik het goed dat mijn ouders beslissen hoe laat ik thuis moet komen na een fuif of een avondje uit.”

- ☐ helemaal akkoord
- ☐ akkoord
- ☐ niet akkoord
- ☐ helemaal niet akkoord

Je geeft de respondenten de kans om te kiezen uit meerdere mogelijkheden, gaande van “ik ben op en top akkoord” tot “ik ben er helemaal tegen”. De schakering in de antwoorden kan je wat duidelijker maken door er een beetje extra uitleg bij te geven. Hierbij moet je erop letten dat die extra uitleg neutraal geformuleerd is, zodat de respondent niet beïnvloed wordt in één of andere richting.

Een mogelijke vorm waarin je jouw enquêtevraag kan opstellen ziet er als volgt uit:

Bewering:

Tot nu toe vind ik het goed dat mijn ouders beslissen hoe laat ik thuis moet komen na een fuif of een avondje uit.

- ☐ **Helemaal akkoord:** Tot nu toe vind ik het goed dat mijn ouders telkens beslissen hoe laat ik thuis moet komen. Ik heb daar geen probleem mee.
- ☐ **Akkoord:** Soms wil ik graag zelf beslissen hoe laat ik terug naar huis kom. Maar als we het niet eens zijn, dan vind ik het meestal toch wel goed dat mijn ouders beslissen.
- ☐ **Niet akkoord:** Ik wil wel met mijn ouders overleggen. Maar als we het niet eens zijn, dan wil ik meestal graag zelf beslissen.
- ☐ **Helemaal niet akkoord:** Ik vind dat ik altijd zelf moet kunnen beslissen wanneer ik terug thuis moet zijn. Mijn ouders zouden daar eigenlijk nooit mogen in tussenkomen.

Je wil graag dat je medeleerlingen “eerlijk” antwoorden op je enquête. Daarom gebruik je best een **anonieme** enquête, waarbij niemand kan achterhalen wie wat geantwoord heeft. Er is dan geen enkele reden om te liegen, en de respondent kan gerust aankruisen wat hij echt denkt.

Voor dit onderzoek ga je één dataset maken die dezelfde is voor de hele klas. Jullie kunnen daarbij als volgt te werk gaan. Vraag aan je leraar de enquêteformulieren en ook een doos om de ingevulde formulieren in te steken. Elke leerling uit de klas krijgt een formulier. Hierop moet iedereen één en slechts één vakje aankruisen. Dan moet iedereen het formulier dichtplooien zodat de anderen niet kunnen zien welk vakje er is aangekruist. Daarna worden alle formulieren in de doos gelegd en de doos wordt eens goed geschud. Op die manier heb je waarschijnlijk al een 15-tal antwoorden. Voor dit onderzoek zou het goed zijn om een 30-tal antwoorden te hebben. Je kan afspreken dat je deze enquête ook mag houden in een andere klas van de tweede graad, of je kan andere leerlingen van jouw leeftijd op de speelplaats vragen om aan die enquête deel te nemen. Zorg dan dat je de nodige formulier en de doos bij de hand hebt. In ieder geval moet je altijd vooraf zeggen dat de enquête **anoniem** is, wat betekent dat het aangekruiste formulier moet dichtgeplooid worden en dat het dan bij de andere formulieren in de doos moet worden gestoken.

- *Voer nu die enquête uit. Volg hierbij de instructies van je leerkracht.*

1.2 Een dataset maken

De informatie uit je enquête zou je, zoals bij de M&M's, kunnen opschrijven in een dataset, met elementen en veranderlijken. Hierbij zijn de elementen de leerlingen die aan deze enquête hebben meegedaan (dat zijn dus de ingevulde enquêteformulieren). Per enquêteformulier kan je de waarde aflezen van de veranderlijke “de mate van akkoord zijn”. Je zou nu, zoals in vorig onderzoek, deze dataset volledig kunnen uitschrijven en voor de 30 elementen vermelden in hoeverre men akkoord is. Maar je kan voor de onderzoeksvraag die hier gesteld wordt ook op een kortere manier te werk gaan. Dat leer je in de volgende paragraaf.

- *Maak een schema voor het volledig opstellen van de dataset, en vul dit in voor een zelf gekozen voorbeeld van 5 antwoorden op de enquête. Geef hierbij goed het verschil aan tussen de naam van de veranderlijke en haar waarden.*

1.3 De dataset: getallen en context

Bij je onderzoek heb je één veranderlijke opgemeten, namelijk “de mate van akkoord zijn”. Die veranderlijke heeft 4 mogelijke waarden: “helemaal akkoord”, “akkoord”, “niet akkoord”, en “helemaal niet akkoord”. Juist zoals bij M&M's zie je dat de veranderlijke niet zo heel veel verschillende waarden heeft, en dat je er geen zinvolle wiskundige bewerkingen kan op uitvoeren. Deze veranderlijke wordt daarom een **kwalitatieve** veranderlijke genoemd.

De situatie die je hier hebt is toch niet helemaal dezelfde als bij M&M's. Kleuren hebben geen volgorde, maar “de mate van akkoord zijn” heeft dat wel. Er is een **orde** te bespeuren. Zo'n veranderlijke wordt daarom **ordinaal** genoemd. Je hebt hier te maken met een **ordinaal kwalitatieve veranderlijke**.

De informatie uit je enquête kan je bondig samenvatten in een frequentietabel, en meer heb je hier niet nodig. Je kan daarbij eerst een schema maken met de geordende uitkomsten in de eerste kolom. Voorzie plaats om te turven (turven = streepjes trekken) in de tweede kolom, en maak een derde kolom voor de frequentie. Kies dan één leerling die formulier per formulier uit de doos haalt en zegt welk antwoord op dat formulier is aangekruist. Alle andere leerlingen volgen mee door op de juiste plaats streepjes te zetten. Op het einde tel je per rij alles samen. Een voorbeeld van zo'n tabel ziet er als volgt uit.

Ouders beslissen over het uur van thuiskomst		
Mate van akkoord	Turven	Frequentie
helemaal akkoord		5
akkoord		12
niet akkoord		8
helemaal niet akkoord		3



Het is belangrijk om voldoende aandacht te schenken aan het soort veranderlijke dat je onderzoekt. Elke soort veranderlijke wordt op een eigen manier behandeld, met een eigen soort van grafieken en tabellen.

- Noteer nu de resultaten van de enquête als een frequentietabel. Schrijf er ook je naam, datum en plaats bij, samen met een korte titel. Voeg ook een origineel enquêteformulier toe.

Mate van akkoord	Turven	Frequentie
helemaal akkoord		
akkoord		
niet akkoord		
helemaal niet akkoord		

2 Op speurtocht in de dataset

2.1 De frequentietabel

De frequentietabel heb je zopas opgesteld.

- *Hoe vind je de steekproefgrootte uit de frequentietabel?*
- *Voeg een kolom toe met de relatieve frequentie. Welke interessante informatie kan je daar rechtstreeks uit aflezen voor dit onderzoek?*

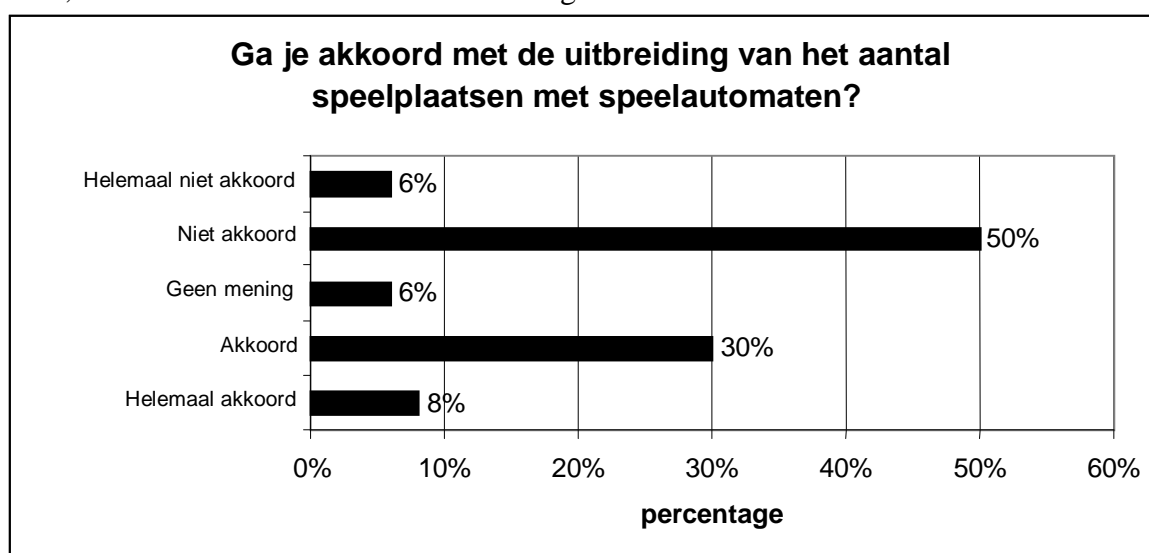
Ouders mogen beslissen over het uur van thuiskomst			
Mate van akkoord	Turven	Frequentie	Relatieve frequentie
helemaal akkoord			
akkoord			
niet akkoord			
helemaal niet akkoord			

2.2 Het staafdiagram

De resultaten van je enquête kan je grafisch voorstellen. Voor een kwalitatieve veranderlijke gebruik je een staafdiagram.

Als je de staven vertikaal plaatst, is het moeilijk om de woordgroepen “helemaal niet akkoord”, “akkoord”, ... op een goede manier onder de gepaste staven te krijgen. De tekst is te lang, en schuin of verticaal geschreven tekst is moeilijk voor de lezer. In het vorige onderzoek heb je gezien hoe je dit probleem kan oplossen. Je tekent de staven gewoon horizontaal. De waarden van de veranderlijke schrijf je onder elkaar op de verticale as. Deze as is dus een **verticale x-as**.

Hier zie je zo’n horizontaal staafdiagram. Het stelt de resultaten voor van een enquête die de gemeente Zwolle in maart 2005 gehouden heeft bij jongeren van die gemeente. Het antwoord “geen mening” moet je daarbij interpreteren als: “het is mij om het even of die uitbreiding er al dan niet komt, ik ben er niet voor en ik ben er niet tegen”.



Zoals je ziet zijn de waarden van de veranderlijke hier niet geordend volgens hun relatieve frequentie.

- Welke ordening kies jij, voor jouw studie, bij dit type staafdiagram (horizontaal)?
- Welk belangrijke verschillen zijn er met het staafdiagram dat bij een nominaal kwalitatieve veranderlijke hoort?

- *Zal jij in je staafdiagram frequenties of relatieve frequenties gebruiken? Waarom?*
- *Maak voor jouw enquête een horizontaal staafdiagram.*



3 Wat heb je gevonden? Hoever kan je gaan in je conclusie?

3.1 De variabiliteit van steekproefresultaten

Met je onderzoek wilde je te weten komen of jongeren van jouw leeftijd het goed vinden dat hun ouders beslissen hoe laat zij thuis moeten zijn. De resultaten die jij hebt gevonden zijn afkomstig van de leerlingen die aan je enquête hebben deelgenomen. Strikt gesproken gelden zij dus enkel voor die groep leerlingen.

- *Is er een duidelijke meerderheid voor één van de vier meningen?*
- *Welke conclusie zou je op basis van jouw cijfermateriaal kunnen formuleren?*

- *Als je nu eens andere leerlingen ondervraagt, kan het dan:*
 - *dat je juist dezelfde resultaten vindt? (wanneer zou dat kunnen gebeuren?)*

 - *dat de nieuwe resultaten goed lijken op wat jij nu hebt? (wanneer zou dat kunnen gebeuren?)*

 - *dat de nieuwe resultaten sterk verschillen van wat je nu hebt? (wanneer zou dat kunnen gebeuren?)*

3.2 Een enquête afnemen

Er is heel wat variatie in de manier waarop enquêtes worden afgenomen. Je kan de respondenten op verschillende manieren benaderen. Hierbij heb je ondermeer:

- de **schriftelijke enquête**, waarbij de respondent een vragenlijst krijgt om in te vullen. Zo'n vragenlijst krijg je meestal per post toegestuurd.
 - de **telefonische enquête**, waarbij je wordt opgebeld door een onderzoeker die je een aantal vragen stelt waarop je moet antwoorden.
 - het **persoonlijk interview**, waarbij de onderzoeker je persoonlijk benadert en vragen stelt. Jouw antwoorden worden door hem op het formulier genoteerd.
-
- *Tot welk type behoort de enquête die jij juist hebt afgenomen?*

 - *Is dit type (in het algemeen) duurder of goedkoper dan de andere types? Waarom?*

- *Verwacht je dat voor dit type meer of minder mensen zullen antwoorden dan voor de andere types? Waarom?*
- *Hoe kan je bij dit type ervoor zorgen dat je een grote respons krijgt? Zal dat altijd mogelijk (of betaalbaar) zijn?*

3.3 Wat kan er mis gaan?

Er kan nogal wat fout lopen bij een enquête. Het onmiddellijke gevolg is dat je de gevonden conclusie niet kan veralgemenen of - erger nog - dat zij helemaal verkeerd is, zelfs voor de groep die jij hebt onderzocht!

3.3.1 Problemen met de vraag

Het is heel moeilijk om een vraag goed te formuleren. Een vraag moet duidelijk zijn en mag niet voor verschillende interpretaties vatbaar zijn. Kleine verschillen in de formulering van de vraag kunnen bij de lezer een andere reactie oproepen.

Voorbeeld

Een identieke vraag op verschillende manieren verwoorden kan aanleiding geven tot een ander antwoord. Psychologen hebben dit reeds uitgetest. Ze gaven aan hun proefpersonen volgende tekst:

Het uitbreken van een nieuwe ziekte heeft 600 mensen dodelijk besmet. Sommigen zouden kunnen gered worden als men er in slaagt om nog snel een nieuw vaccin te ontwikkelen. Dit kan op twee manieren gebeuren, maar de tijd dringt en men moet kiezen voor de ontwikkeling van ofwel vaccin A ofwel vaccin B.

Aan een eerste groep respondenten werd gevraagd aan welk vaccin zij de voorkeur gaven. Zij kregen hierbij de volgende informatie:

- als men kiest voor vaccin A dan zullen 200 mensen gered worden
- als men kiest voor vaccin B dan is er $\frac{1}{3}$ kans dat alle 600 mensen gered worden en $\frac{2}{3}$ kans dat niemand zal gered worden.

Ook aan een tweede groep respondenten werd gevraagd welk vaccin zij zouden verkiezen, maar hen werd volgende informatie verstrekt:

- als men kiest voor vaccin A dan zullen 400 mensen sterven
- als men kiest voor vaccin B dan is er $\frac{1}{3}$ kans dat niemand zal sterven, en $\frac{2}{3}$ kans dat alle 600 mensen zullen sterven.

Je merkt dat de informatie voor beide groepen identiek was, alleen anders verwoord. De 2 groepen respondenten reageerden nochtans sterk verschillend.

- *Welk verschil bemerk je tussen de vraag die je je leeftijdsgenoten hebt voorgelegd en de onderstaande tekst? Zou dit bij sommige leerlingen tot een ander antwoord kunnen leiden? Leg uit waarom.*
 - *Ik vind het goed dat het de ouders zijn die, bij jongeren van 14 – 15 jaar, beslissen hoe laat zij moeten thuiskomen na een fuif of een avondje uit.*

3.3.2 Problemen met de respons

De **respons** zijn de mensen die antwoorden op de enquête. De mensen die wel gevraagd zijn om aan de enquête deel te nemen, maar die hun formulier niet terugsturen (of die weigeren mee te werken) noem je de **non-respons**. Respons en non-respons worden meestal uitgedrukt in percenten. Non-respons kan de waarde van een onderzoek verminderen, of zelfs helemaal teniet doen.

Voorbeeld.

Een leerkracht wou weten of de leerlingen op haar school het fijn vonden om enquêtes in te vullen. Om dat te onderzoeken ontwierp zij een enquête en stuurde die naar alle leerlingen van de school. Van de formulieren die zij terugkreeg was de overgrote meerderheid positief. Daarop besloot die leerkracht dat de leerlingen van haar school positief stonden tegenover het invullen van enquêtes.

- *Wat denk jij over dit voorbeeld? Zou er een verband kunnen zijn tussen je houding tegenover het invullen van enquêtes en het feit of je daar al dan niet aan meedoet? Wat zou je bijkomend moeten weten om het besluit van de leerkracht te kunnen bevestigen?*

3.3.3 Problemen met de selectie van de deelnemers

Het is niet altijd mogelijk om een EAS (enkelvoudige aselechte steekproef) te trekken. Er zijn heel wat andere manieren om deelnemers te selecteren. Maar je moet er wel altijd op letten dat je selectiemethode een goede representatie van de populatie toelaat. Als je een methode gebruikt die bijvoorbeeld bepaalde groepen systematisch uitsluit, dan zit je verkeerd.

Voorbeeld.

In 1936 liep de eerste termijn van het presidentschap van F.D. Roosevelt ten einde en waren er opnieuw verkiezingen. De tegenkandidaat van (democraat) Roosevelt was de republikeinse gouverneur A. Landon uit Texas. De meeste waarnemers dachten dat Roosevelt voor een tweede ambtstermijn zou herkozen worden, maar de enquête van het magazine “*Literary Digest*” voorspelde

iets helemaal anders. Op basis van de 2.4 miljoen antwoorden die waren binnengekomen (het grootste aantal mensen in de geschiedenis dat ooit op een enquête heeft geantwoord) voorspelden zij dat Landon zou winnen met 57% van de stemmen tegenover slechts 43% voor Roosevelt. Maar de verkiezingen draaiden uit op een klinkende overwinning voor Roosevelt, met 62% tegenover 38%.

Hoe kon de “*Literary Digest*” (die kort nadien failliet ging) op basis van zo’n reuzengrote enquête toch nog die enorme fout maken? Het antwoord zit zowel in de non-respons als in de **vertekende selectie** van de deelnemers. De “*Literary Digest*” had namelijk 10 miljoen formulieren per post verstuurd en had daarvoor adressen gebruikt van hun lezerslijst maar ook uit telefoonboeken enz.. In die tijd had $\frac{3}{4}$ van de mensen geen telefoon, en was de armere groep van de bevolking zeker niet geabonneerd op de “*Literary Digest*”. Er werden dus bepaalde groepen van mogelijke kiezers systematisch uitgesloten. Verder was er slechts 20% respons. De conclusie van deze enquête was dus totaal waardeloos.



Als je een vertekende methode gebruikt om deelnemers te selecteren, dan helpt het niet om een grote steekproef te trekken. Je herhaalt dan alleen maar een essentiële fout op een veel grotere schaal.

- *Als je aan de ingang van een supermarkt totaal willekeurig mensen aanspreekt om enkele vragen te beantwoorden, werk je dan met een EAS?*
- *Hoe heb jij de deelnemers gekozen voor je eigen enquête? Hoever kan je gaan als je de resultaten zou willen veralgemenen? Voor welke populatie zou je dat dan willen doen? Hoe zou je de deelnemers dan moeten selecteren?*

4 Kernachtige samenvatting van dit onderzoek

Je samenvatting bestaat opnieuw uit twee delen:

- *De antwoorden op de contextvragen (de www-vragen)*
- *De besluiten over het uitgevoerde onderzoek*



Herinner je dat de besluiten van een statistisch onderzoek maar betekenis krijgen als je ook de achtergrond van het onderzoek kent.
In veel krantenartikels springt men daar nogal lichtzinnig mee om!

- *Formuleer in een bondige tekst de antwoorden op de contextvragen.*
- *Formuleer in een bondige tekst je besluiten over het uitgevoerde onderzoek.*

5 Zelfevaluatie

In dit onderzoek heb je geleerd over:

- de enquête
- de soorten vragen bij een enquête
- de manieren om een enquête af te nemen
- de vertekening door de vraagstelling
- de vertekening door de non-respons
- de vertekening door de selectie
- het turven van categorische gegevens
- de ordinaal kwalitatieve veranderlijke
- het staafdiagram bij een ordinaal kwalitatieve veranderlijke

Je bent nu in staat om de volgende opdrachten uit te voeren:

- *Zeg kort in eigen woorden wat een enquête is, en op welke manier je daarbij mensen kan contacteren.*
- *Kan de manier waarop een vraag geformuleerd is een invloed hebben op het antwoord? Kan je hierbij zelf een voorbeeld bedenken?*
- *Wanneer noem je een veranderlijke ordinaal kwalitatief? Geef een (nieuw) voorbeeld van een ordinaal kwalitatieve veranderlijke.*

- *Welke verschillen zijn er tussen een staafdiagram bij een ordinaal kwalitatieve veranderlijke en een staafdiagram bij een nominaal kwalitatieve veranderlijke?*
- *Kan je problemen aangeven waardoor een enquête op het Internet waardeloos zou kunnen zijn?*
- *Is het een goed idee om op zaterdagvoormiddag toevallige personen aan te spreken in de winkelstraat als je wil weten wat de inwoners van die stad denken over de werking van de gemeenteraad?*
- *Onderstaand fragment is afkomstig van het jongerenonderzoek 2001-2002 Euregio Maas-Rijn, uitgevoerd door de provincie Limburg in samenwerking met het Centrum voor Statistiek van de Universiteit Hasselt. Welke soort vragen bemerk je hier? Leg uit. Welk type veranderlijke wordt er opgemeten bij de eerste vraag? Wat zijn haar waarden?*

ROKEN	
Heb je wel eens sigaretten gerookt, ook al was het maar één sigaret of een paar trekjes?	<input type="checkbox"/> neen, ik heb nooit gerookt <input type="checkbox"/> ja, ik heb 1 of 2 keer gerookt <input type="checkbox"/> ja, ik heb vroeger gerookt maar ben nu gestopt <input type="checkbox"/> ja, ik rook af en toe maar niet elke dag <input type="checkbox"/> ja, ik rook elke dag
Op een dag dat je rookt, hoeveel sigaretten rook je dan?	Ik rook dan ongeveer <input type="text"/> <input type="text"/> sigaretten per dag.

- Lees het artikel “Internetpeiling: ‘Belgacom-televisie slaat niet aan’ ” uit De Morgen van 27 juli 2005. Enig idee “bij wie” die digitale televisie niet aanslaat? Is het hier de bedoeling om iets te weten over een grotere populatie? Welke zou dat dan wel zijn? Wordt daar iets over gezegd? Is een internetpeiling een goede methode om een eigenschap van een populatie te onderzoeken? Noem enkele mogelijke problemen.

Internetpeiling: ‘Belgacom-televisie slaat niet aan’

Volgens een peiling op de website digitaletelevisiewijzer.be kan telecomoperator Belgacom maar op weinig interesse rekenen voor zijn pas gelanceerde aanbod aan digitale televisie. Meer dan 3.800 surfers vulden via de website vragen in over welke operator ze willen kiezen voor digitale televisie. Belgacom wist maar 22 procent van de bezoekers van de website te bekoren, terwijl erfconcurrent Telenet 52 procent haalde; 26 procent van

de surfers is nog onbeslist.

Bovendien zou de consument niet erg geneigd zijn veel extra te betalen voor digitale televisie. Tweeëntwintig procent van de respondenten in een andere bevraging op dezelfde website geeft aan niet bereid te zijn extra te betalen voor digitale televisie. “Zolang er buiten de extra televisiekanalen en het voetbalaanbod niet veel extra toepassingen worden aangeboden op digitale televisie, zijn mensen blijkbaar niet erg ge-

neigd er extra voor te betalen”, zegt Sven Bergiers, een informaticastu-

Mensen zijn niet geneigd extra te betalen voor digitale televisie

dent aan de Katholieke Hogeschool Leuven, en een van de mensen achter digitaletelevisiewijzer.be.

Toch liggen ze bij Belgacom niet erg wakker van de cijfers. Volgens woordvoerder Jan Margot heeft eigen marktonderzoek uitgewezen dat er wel degelijk veel potentieel zit in het digitale televisieaanbod van Belgacom. “Je kunt je ook afvragen hoe betrouwbaar de enquête is”, luidt het nog. Belgacom wil nog geen cijfers bekendmaken over de verkoop van digitale-televisieabonnementen, maar volgens Margot verloopt de lancering naar wens.

Een statistisch onderzoek naar honden en katten in een gezin

Zelf een steekproef trekken

In je eerste onderzoek was het zakje M&M's een steekproef uit de populatie van alle M&M's. Maar die steekproef had je eigenlijk niet zelf getrokken. Je had ze gewoon gekregen.

Bij je tweede onderzoek heb je zomaar leerlingen gekozen op een manier die jou het beste uitkwam. Dat is geen goede methode als je daarna een algemene uitspraak wil doen.

Bij de volgende twee onderzoeken wil je telkens iets te weten komen over alle leerlingen van je school zonder daarom al die leerlingen te moeten ondervragen. Je zal hiervoor op een professionele manier een enkelvoudige aselecte steekproef trekken. Dit vraagt inspanning en tijd. Doe het toch maar. Zo ontdek je zelf wat er bij een goed statistisch onderzoek allemaal komt kijken. Maak goede afspraken met je leerkracht zodat alles efficiënt verloopt.

Om te starten moet je weten wat de populatie is. Daarom heb je een lijst nodig met de namen van alle leerlingen van heel je school. Naast elke naam plaats je een volgnummer, bijvoorbeeld van 1 tot 512 als er in je school 512 leerlingen zijn. Nu ga je een lukrake steekproef van grootte 40 trekken. Om dat goed te doen gebruik je het toeval waarbij je de toevalsgenerator in je GRM het werk laat doen.

Zorg ervoor dat het programma TREKZNDR in je GRM staat. Druk nu **[PRGM]**, kies TREKZNDR en druk **[ENTER]**. Beantwoord de vragen die het programma stelt. Het eerste getal dat je moet ingeven is de grootte van de totale populatie. Dat is bijvoorbeeld 512. Daarna wordt gevraagd hoe groot de steekproef moet zijn. Hier tik je 40. Als resultaat krijg je nu 40 toevallige getallen in **[L1]** en de namen die daarbij horen zijn de 40 leerlingen van jouw school die jij zal aanspreken voor het derde en vierde onderzoek. Je kan nu je klas indelen in 5 groepjes die elk 8 leerlingen gaan ondervragen. Daarna leg je alle resultaten samen zodat je één steekproef hebt van grootte 40, die je samen bestudeert. Druk **[STAT]** en 1:Edit... om de lijst **[L1]** te bekijken. Het programma heeft deze lijst automatisch gesorteerd.



<pre> 3:00 EDIT NEW 1:DISCRTBL 2:HISDICH1 3:STAFDGR 4:TREKZNDR </pre>	<p>Grootte van de totale populatie</p> <p>T=</p>	<p>Steekproef : hoeveel trekken zonder terugl.?</p> <p>N=40</p>																																
<p>Resultaat staat in de lijst L1</p> <p>Druk op ENTER</p>	<pre> 3:00 CALC TESTS 1:Edit... 2:SortA(3:SortD(4:ClrList 5:SetUpEditor </pre>	<table border="1"> <thead> <tr> <th>L1</th><th>L2</th><th>L3</th><th>1</th></tr> </thead> <tbody> <tr><td>438</td><td></td><td></td><td></td></tr> <tr><td>453</td><td></td><td></td><td></td></tr> <tr><td>466</td><td></td><td></td><td></td></tr> <tr><td>487</td><td></td><td></td><td></td></tr> <tr><td>488</td><td></td><td></td><td></td></tr> <tr><td>503</td><td></td><td></td><td></td></tr> <tr><td>...</td><td></td><td></td><td></td></tr> </tbody> </table> <p>L1(40) =</p>	L1	L2	L3	1	438				453				466				487				488				503				...			
L1	L2	L3	1																															
438																																		
453																																		
466																																		
487																																		
488																																		
503																																		
...																																		

- Vraag nu verdere instructies aan je leerkracht... en ga op stap.

1 Wat wil je weten? Hoe ga je meten?

1.1 De onderzoeksvraag

Heb je enig idee hoe het zit met de huisdieren van je vrienden en vriendinnen? Weet je welke huisdieren zij allemaal hebben?

Misschien begin je best met alleen maar naar honden en katten te kijken. Over deze dieren kan je heel wat willen weten: het gewicht, de kleur van de pels, enz. Hou het hier maar eenvoudig en tel gewoon hoeveel honden en hoeveel katten er per gezin zijn. Dat wil je weten voor alle gezinnen die kinderen hebben die bij jou op school zitten. Je gaat die natuurlijk niet allemaal ondervragen, en daarom werk je met een steekproef.

- *Welke populatie wordt er hier onderzocht en welke vragen worden er over deze populatie gesteld?*

Je hebt zopas een enkelvoudige aselechte steekproef van grootte 40 getrokken uit de populatie van alle leerlingen van je school.

- *Kan je de steekproef die je pas getrokken hebt gebruiken om dit onderzoek uit te voeren? Wat zou een probleem kunnen zijn en hoe ga je dat oplossen? Wat is je dataset hier?*

1.2 De dataset: getallen en context



Eén van de veranderlijken die je per gezin hebt opgemeten is het aantal honden. Deze veranderlijke heeft waarschijnlijk niet veel verschillende waarden (nul, één, twee, drie, en misschien ook nog vier of vijf). Bovendien is “aantal” een geheel getal.

Het aantal honden is een voorbeeld van een “**discreet numerieke**” veranderlijke.

De naam “**numeriek**” zegt dat het echt over getallen gaat waarmee je kan rekenen. De naam “**discreet**” wijst erop dat de uitkomsten uit elkaar liggen. In dit voorbeeld springen de mogelijke uitkomsten vooruit met één eenheid (tussen 2 en 3 ligt bijvoorbeeld 2.6, maar 2.6 honden krijg je nooit als antwoord!).

2 Op speurtocht in de dataset

Omdat, zoals in de meeste onderzoeken, de oorspronkelijke getallen niet erg overzichtelijk zijn, ga je die samenvatten in een tabel en een grafiek.

Neem je GRM en herstel (indien nodig) de standaardlijsten.

Als de lijsten [L5] en [L6] vol getallen zouden staan dan maak je die als volgt snel leeg. Druk [STAT] en dan 1:Edit... Loop met de pijltjes naar de lijst [L5] en ga helemaal op de kop staan (dus op de naam zelf). Druk dan [CLEAR] en [ENTER]. Indien nodig doe je hetzelfde met de lijst [L6].

L4	L5	L6	5
-----	22 56 17 2 32 99 51	-----	
L5 = {22, 56, 17, 2, ...}			

L4	L5	L6	5
-----	-----	-----	
L5(1)=			

Tik nu je data in je GRM. In lijst [L5] tik je het aantal honden per gezin en in lijst [L6] zet je het bijhorende aantal katten van dat gezin. Kijk na het inbrengen van de gegevens alles nog eens grondig na!

L4	L5	L6	6
	1 2 1 1 3 1 1 0	2 0 0 0 0 4 1 0	
L6(36)=0			

2.1 Een frequentietabel opstellen

Begin met de studie van het aantal honden per gezin.

Je kan weer een tabel maken met drie kolommen. In de eerste kolom zet je het aantal honden per gezin, in de tweede kolom schrijf je hoeveel gezinnen er zijn die dit aantal honden hebben (de frequentie) en in de derde kolom komt de relatieve frequentie.



Dit lijkt goed op de frequentietabellen die je maakte bij kwalitatieve veranderlijken. Toch is er een belangrijk verschilpunt. In de eerste kolom plaats je alle mogelijke uitkomsten, vanaf je kleinste opmeting tot je grootste. Je moet dus alle “mogelijke” tussenliggende uitkomsten opschrijven, ook als bijvoorbeeld “drie honden” in jouw onderzoek niet voorkwam. Geef dan aan die 3 een frequentie nul in de tweede kolom.

Bij discreet numerieke veranderlijken kan je je handig laten helpen door je GRM.

- Zorg dat het programma FREQDISC in je GRM staat.
- Kopieer eerst de lijst met het aantal honden (lijst [L5]) naar lijst [L1]. Druk **[2nd]** **[L5]**, dan op **[STO→]** om het pijltje te maken en vervolgens op **[2nd]** **[L1]**. Druk daarna op **[ENTER]** om het commando uit te voeren. Je ziet dan het begin van de lijst getallen die in [L1] zijn terechtgekomen. In [L5] verandert niets.
- Druk **[PRGM]** zodat je de lijst krijgt van alle programma's in je GRM. In dit voorbeeld zie je dat het programma FREQDISC naast het nummer 4 staat en daarom moet je hier op 4 drukken. In jouw toestel kan dat een ander nummer zijn. Druk op het juiste nummer. Op je scherm verschijnt het commando prgmFREQDISC. Druk op **[ENTER]** om dit programma te laten lopen.
- Na een korte tijd zie je een melding dat de waarden in de lijst [L2] staan, de frequenties in [L3] en de relatieve frequenties in [L4].
- Bekijk nu wat er in die lijsten staat. Druk **[STAT]** en dan 1:Edit.
- Je ziet dat er, in dit onderzoek, 18 gezinnen zijn zonder hond, 14 gezinnen met één hond, 3 gezinnen met twee honden en 1 gezin met drie honden.
- In [L4] kan je de relatieve frequenties aflezen.

<pre>L5→L1 (0 0 1 2 0 0 0 ...</pre>																								
<pre>PRGM EDIT NEW 1:CLS2X 2:CLSVARS 3:FREQCONT 4:FREQDISC 5:HISDICH 6:HYPGEOM 7↓INTERVAL</pre>																								
<pre>L5→L1 (0 0 1 2 0 0 0 ... PRGMFREQDISC</pre>																								
Done																								
<table border="1" style="width: 100%; border-collapse: collapse;"><thead><tr><th style="width: 33%;">L2</th><th style="width: 33%;">L3</th><th style="width: 33%;">L4</th><th style="width: 33%;">2</th></tr></thead><tbody><tr><td>0</td><td>18</td><td>.5</td><td></td></tr><tr><td>1</td><td>14</td><td>.38889</td><td></td></tr><tr><td>2</td><td>3</td><td>.08333</td><td></td></tr><tr><td>3</td><td>1</td><td>.02778</td><td></td></tr><tr><td colspan="4" style="text-align: center;">-----</td></tr></tbody></table>	L2	L3	L4	2	0	18	.5		1	14	.38889		2	3	.08333		3	1	.02778		-----			
L2	L3	L4	2																					
0	18	.5																						
1	14	.38889																						
2	3	.08333																						
3	1	.02778																						

L2(5) =																								

- *Maak een frequentietabel voor het aantal honden per gezin. Voeg ook een kolom met de relatieve frequentie toe.*

Aantal honden per gezin	Frequentie = hoeveel gezinnen met dit aantal honden	Relatieve frequentie

- *Maak de som van de getallen in de kolom met de frequenties. Hoeveel is dat? Waarom?*
- *Maak de som van de getallen in de kolom met de relatieve frequenties. Hoeveel is dat? Waarom?*

2.2 Een staafdiagram tekenen

Het staafdiagram is de basisfiguur voor een discreet numerieke veranderlijke met een beperkt aantal verschillende uitkomsten. Het “aantal honden per gezin” is zo’n veranderlijke en dus teken je daarvoor een staafdiagram.

Op de x-as duid je alle mogelijke uitkomsten aan die je had kunnen vinden, vanaf je kleinste tot je grootste observatiegetal. Dit is bijvoorbeeld 0, 1, 2, en 3. In de y-richting teken je boven elk van die mogelijke uitkomsten een staafje. Als je de lengte van dat staafje gelijk neemt aan de frequentie, dan heb je een grafische voorstelling van de eerste twee kolommen van je frequentietabel.



Bij de vorige onderzoeken gebruikte je ook het staafdiagram bij kwalitatieve veranderlijken. Maar de voorstellingswijze voor nominaal en ordinaal was niet dezelfde.

Nu merk je terug een verschil. De waarden op de x-as zijn discreet numeriek. Je mag hier net zoals bij de frequentietabel geen waarden overslaan! Omdat de mogelijke uitkomsten uit elkaar liggen, zullen de staafjes niet tegen elkaar getekend worden.

Laat je helpen door je GRM om een staafdiagram voor het aantal honden per gezin te tekenen.

- Zorg dat het programma STAAFDGR in je GRM staat.
- Druk **[PRGM]**, kies STAAFDGR en druk **[ENTER]**. Met de pijltjes kan je de figuur doorlopen. Druk meerdere keren op **[▶]** of op **[◀]** om eenzelfde staafje te doorlopen en kijk goed waar de cursor staat. Onderaan zie je telkens de waarde van x en y. Om het programma te stoppen druk je nog eens op **[ENTER]**.

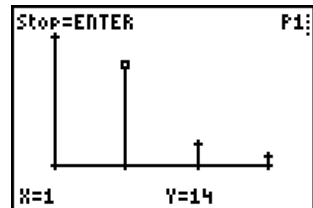
```

EDIT NEW
1:DISCRTEL
2:HISDICH1
3:STAAFDGR
4:TREKZNDR
  
```

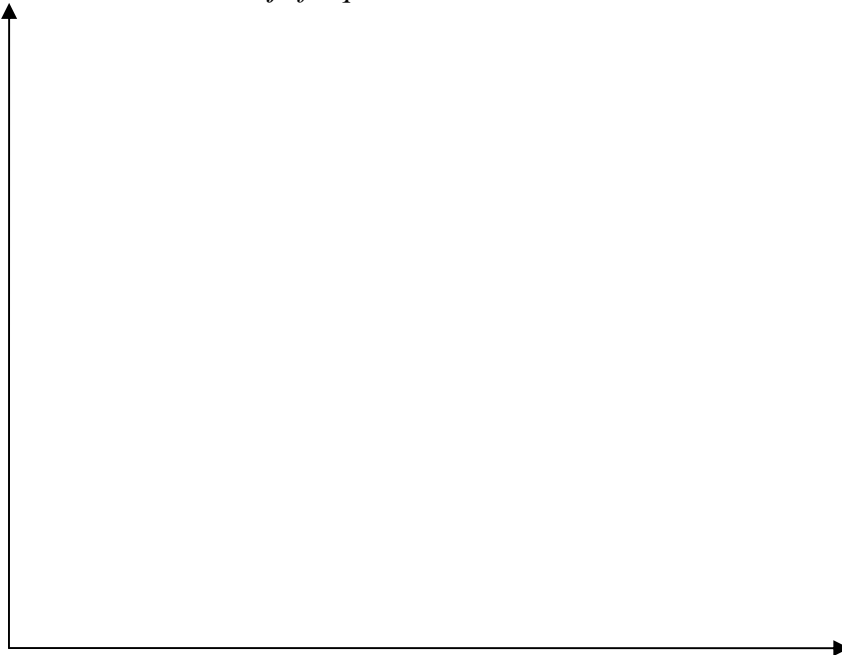
```

Bekijk de figuur
Trace is actief

Ga verder= ENTER
  
```



- *Teken een staafdiagram voor het aantal honden per gezin. Voorzie de assen van de juiste naam. Als je de lengte van de staafjes gelijk neemt aan de frequentie, dan heb je een grafische voorstelling van de eerste twee kolommen van je frequentietabel.*

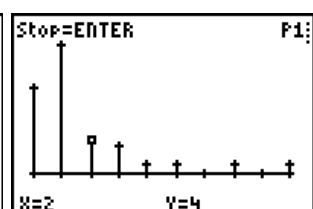


Herhaal nu alle vorige bewerkingen voor huisdieren per gezin (honden en katten). Om te starten, plaats je de som van de lijsten [L5] en [L6] in lijst [L1].

```

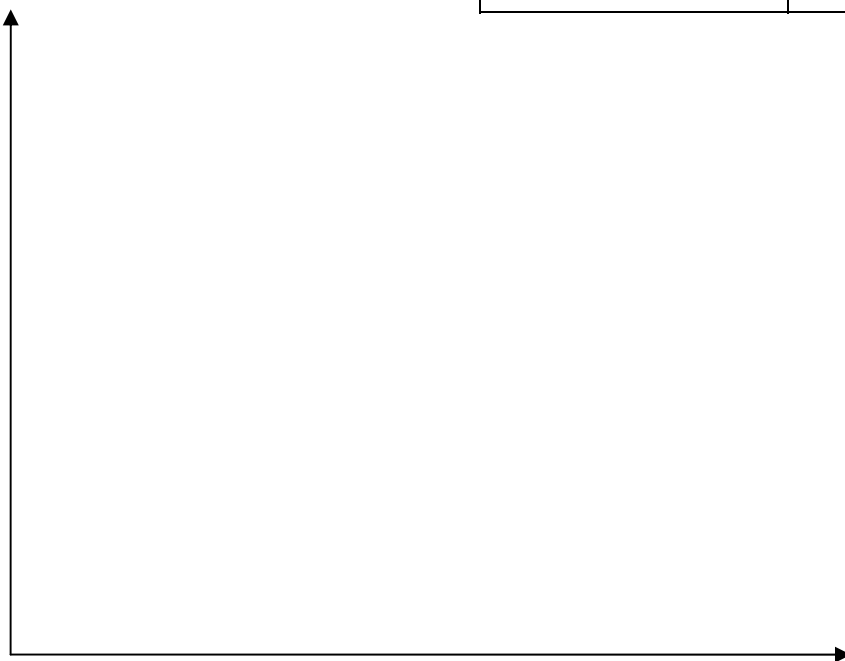
L5+L6→L1
{3 0 4 9 1 1 1 ...
PRGMFREQDISC
  
```

L1	L2	L3	1
3	0	10	
0	1	15	
4	2	4	
9	3	3	
1	4	1	
1	5	1	
1	6	0	
L1(7)=1			



- *Stel de frequentietabel op voor huisdieren (= honden en katten) per gezin en teken het staafdiagram. Laat je helpen door de GRM.*

<i>Aantal huisdieren per gezin</i>	<i>Frequentie = hoeveel gezinnen met dit aantal huisdieren</i>




2.3 Numerieke kenmerken: gemiddelde en mediaan

Hoe het met het totale aantal huisdieren (honden en katten) gesteld is bij die 36 gezinnen wordt mooi weergegeven in je staafdiagram. Maar wat doe je als men vraagt om al die informatie in enkele getallen samen te vatten?

Een samenvatting in getallen (die “kengetallen” worden genoemd) geeft je zelden evenveel informatie als een goede figuur. Maar soms kan een kengetal dienen als “typisch” resultaat. Dat is natuurlijk wel handig.

Als eerste kenmerk wil je weten hoeveel huisdieren een “typisch” gezin van je onderzoek heeft. Daarom ga je op zoek naar een getal dat “het centrum” van al je resultaten weergeeft. Een gebruikelijk kengetal hiervoor is **het gemiddelde**. Een ander kengetal voor “centrum” is **de mediaan**.

 *Lees nu in je infoboekje de bijkomende informatie over gemiddelde en mediaan bij een discreet numerieke veranderlijke.*

Het gemiddelde en de mediaan laat je berekenen door je GRM. Het aantal huisdieren per gezin staat nog altijd in [L1]. Op deze lijst kan je bewerkingen uitvoeren. Druk op **[2nd]** **[LIST]**.

- Loop met de cursor naar MATH, loop dan naar beneden naar 3:mean(en druk op **[ENTER]**.
- Vervolledig nu het commando op je scherm door op **[2nd]** **[L1]** te drukken, gevolgd door **[ENTER]**. Het resultaat is het gemiddelde van de getallen in lijst [L1]. In dit voorbeeld is dat na afronding 1.6.
- Op een volledig analoge manier bepaal je de mediaan van de getallen in de lijst [L1].

NAME	OPS	MATH	NAME	OPS	MATH
1:	L1		1:	min(
2:	L2		2:	max(
3:	L3		3:	mean(
4:	L4		4:	median(
5:	L5		5:	sum(
6:	L6		6:	prod(
7:	TEMP1		7:	stdDev(
mean(L1)			mean(L1)		
			1.583333333		
			median(L1)		
			1		



Bij de vorige onderzoeken hebben we het niet gehad over kengetallen, zoals gemiddelde en mediaan. Dat kon natuurlijk niet, want met kleuren bijvoorbeeld kan je niet rekenen. Als je met numerieke veranderlijken werkt, kan dat wel. Zeg ook waarover het kengetal gaat, in de context van je onderzoek. Gebruik daarbij de juiste eenheid, zoals: een gemiddelde van 1.6 “huisdieren per gezin”.

- *Hoeveel huisdieren zijn er gemiddeld per gezin? Gebruik je GRM en let op de juiste notatie.*
- *Wat is de mediaan van het aantal huisdieren per gezin? Gebruik je GRM en let op de juiste notatie.*

2.4 Een staafdiagram interpretern

Probeer uit een grafiek zoveel mogelijk informatie af te lezen. Leer kijken naar “globale” vormen. Zoek naar “grote patronen” maar ook naar onderbreking van patronen, zoals eigenaardige gaten, pieken of clusters (ophopingen). Probeer daarvoor een zinvolle verklaring te geven. Breng ook het gemiddelde en de mediaan in verband met de grafiek.



Lees nu in je infoboekje de bijkomende informatie over staafdiagrammen bij een discreet numerieke veranderlijke.

Bekijk de globale kenmerken van je staafdiagram voor het aantal huisdieren per gezin.

- *Heb je een symmetrische figuur of is de figuur scheef (en naar welke kant)?*
- *Heb je voor deze vorm een zinvolle uitleg?*
- *Zijn er eigenaardige gaten, of clusters, of pieken te bespeuren? Zijn die te verklaren door het toeval van je opmetingen of heb je een andere zinvolle uitleg?*
- *Had je op basis van je staafdiagram vooraf kunnen zeggen welk kengetal het grootste zou zijn, het gemiddelde of de mediaan? Waarom?*
- *Stemt dat overeen met het gemiddelde en de mediaan die je gevonden hebt?*

3 Wat heb je gevonden? Hoever kan je gaan in je conclusie?

3.1 De variabiliteit van steekproefresultaten

Voor jouw steekproef heb je een staafdiagram getekend voor het aantal huisdieren per gezin en je hebt ook het gemiddelde en de mediaan berekend. Dat zijn jouw resultaten.

- *Als volgende week een andere klas uit je school hetzelfde onderzoek op dezelfde manier uitvoert, zal die dan hetzelfde staafdiagram, hetzelfde gemiddelde en dezelfde mediaan vinden?*
- *Kan je je antwoord op vorige vraag wat verduidelijken door te beschrijven hoe die andere klas de steekproef trekt. Is hun methode om de steekproef te trekken verschillend van de methode die jouw klas gebruikt? Geeft “dezelfde methode” ook “dezelfde uitkomsten”?*

3.2 Een uitspraak over de populatie

Het woord zegt het zelf: in de “exploratieve” statistiek ga je op “exploratie” in je dataset. Jij hebt dat voor dit onderzoek gedaan. Voor het aantal huisdieren bijvoorbeeld ken je nu het gemiddelde en de mediaan. Je hebt ook ontdekt dat je staafdiagram scheef naar rechts is en je hebt dat op een zinvolle manier proberen te verklaren.

Conclusies van een “exploratief” onderzoek zijn in de eerste plaats van toepassing op de dataset die jij hebt onderzocht, en dus op de elementen die daarin voorkomen (de door jou geselecteerde gezinnen). Met goede statistische methoden kunnen die conclusies veralgemeend worden. Je krijgt dan geen exacte uitspraken over de totale populatie maar goede benaderingen waarvan je de betrouwbaarheid kan aangeven. Bij dit alles is de manier waarop je een onderzoek uitvoert (zoals het trekken van de steekproef) van cruciaal belang.

Wat kan je nu zeggen over het aantal huisdieren bij alle gezinnen die een kind op je school hebben?

- *In je steekproef vond je dat er gemiddeld 1.6 huisdieren per gezin waren. Nu vraagt men wat het gemiddeld aantal huisdieren per gezin is in de hele populatie. Jij zegt dat dit 1.6 is. Hoe kan je dit antwoord beter formuleren?*

- *In die andere klas vonden ze een gemiddeld aantal huisdieren per gezin dat gelijk was aan 1.5, en zij besluiten daaruit dat er gemiddeld 1.5 huisdieren per gezin zijn in de hele populatie. Wie heeft er nu eigenlijk gelijk? Of gaat het niet over “gelijk hebben”?*

Een uitgewerkt “statistisch” antwoord op bovenstaande vraag krijg je in de derde graad. Maar je hoeft niet zolang te wachten om nu al je gezond verstand te gebruiken en hier iets zinvol over te zeggen.

4 Kernachtige samenvatting van dit onderzoek

Je samenvatting bestaat opnieuw uit twee delen

- *De antwoorden op de contextvragen (de www-vragen)*
- *De besluiten over het uitgevoerde onderzoek*

Omdat je met numerieke gegevens werkt, kan je hier de centrummaten (gemiddelde en mediaan) vermelden, en tevens zeggen hoe zinvol ze zijn voor jouw onderzoek. Vergeet ook nooit om een goede grafiek te tekenen en die te interpreteren.

- *Formuleer nu de antwoorden op de contextvragen (de www-vragen)*

- *Formuleer de besluiten over het uitgevoerde onderzoek.*

5 Zelfevaluatie

In dit onderzoek heb je geleerd over:

- de discreet numerieke veranderlijke
- de frequentietabel bij een discreet numerieke veranderlijke
- het staafdiagram bij een discreet numerieke veranderlijke
- het gemiddelde en de mediaan als centrummaten
- de interpretatie van centrummaten in combinatie met een staafdiagram.

Je bent nu in staat om volgende opdrachten uit te voeren:

- *Wanneer is een veranderlijke “numeriek”? Geef een voorbeeld van een numerieke veranderlijke en geef ook een voorbeeld van een veranderlijke die niet numeriek is. Is een getal altijd te beschouwen als een numerieke veranderlijke?*
- *Wanneer is een numerieke veranderlijke discreet? Leg dat uit in je eigen woorden. Is het aantal kinderen per gezin een discreet numerieke veranderlijke? Waarom?*
- *Waar moet je speciaal op letten als je een frequentietabel van een discreet numerieke veranderlijke opstelt? Heeft dat gevolgen voor het bijhorende staafdiagram?*
- *Geef een voorbeeld van een vraag (in de context van het onderzoek naar het aantal honden) waarbij je antwoordt met frequenties en niet met relatieve frequenties.*

- *Geef ook een voorbeeld van een vraag in dezelfde context waarbij je beter met relatieve frequenties werkt.*
- *Welke eigenschap probeert het gemiddelde te beschrijven? Soms lukt dit goed maar soms ook niet. Hoe kan je dat zien op een staafdiagram? Zeg in woorden hoe je het gemiddelde berekent. Kun je daaruit afleiden of het gemiddelde gevoelig is voor uitschieters? Kan je daarvan een eenvoudig voorbeeld geven?*
- *Welke eigenschap probeert de mediaan te beschrijven? Soms lukt dit goed maar soms ook niet. Hoe kan je dat zien op een staafdiagram (verwijs naar een voorbeeld in de infotekst)? Zeg in woorden hoe je de mediaan berekent. Kan je daaruit afleiden of de mediaan gevoelig is voor uitschieters? Kan je daarvan een eenvoudig voorbeeld geven?*

- 300 personen hebben hun naamkaartje in een doos gelegd. Jij moet hieruit 14 namen trekken die een gratis weekend aan zee krijgen. Zeg eerst hoe jij dat zou doen met de doos en de kaartjes. Zeg daarna hoe je dit beter kan doen, en gebruik daarbij de uitdrukking “enkelvoudige aselechte steekproef”. Werk met je GRM, leg uit wat je doet, en zeg wat je daarna nog moet doen om de namen van de winnaars te kennen.
- Lees het artikel “Vaders ontbijten slechtst” uit Het Belang van Limburg van 14 januari 2005. Over welke populatie van gezinnen zou het hier gaan? Wordt er gezegd op welke manier de ondervraagde gezinnen geselecteerd werden? Is er uit de populatie een enkelvoudige aselechte steekproef getrokken? Kan dit een invloed hebben op de eindconclusie? Wat zegt het artikel hier zelf over?

Vaders ontbijten slechtst

Vier op de tien gezinnen zit 's morgens samen aan tafel

HASSELT - Vaders hebben, binnen het gezin, de slechtste ontbijtgewoonten. En kinderen slaan veel minder hun ontbijt over dan gedacht. Iedereen weet zo ongeveer wel wat een gezond ontbijt inhoudt, maar daarom ontbijt men nog niet gezond. Dat zijn een paar resultaten van een grootschalig ontbijtonderzoek, uitgevoerd door de Gezinsbond.

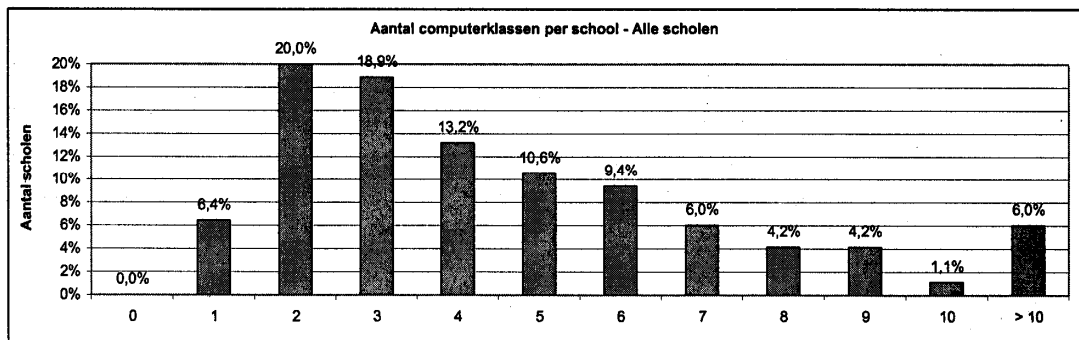
Al vier jaar lang probeert de Gezinsbond families van de noodzaak van een gezond ontbijt te overtuigen. 60.000 Vlamingen werden uitgenodigd rond de ontbijt-tafel. De Gezinsbond stelde een vragenlijst samen die door 6.989 leden werd ingevuld en die betrekking had op 28.000 Vlamingen, vier per gezin.

Alles bij elkaar valt het nog mee. Minder dan twee op de honderd kinderen vertrekt systematisch met lege maag naar school. De Gezinsbond is daar bijzonder blij om, maar relativeert. De leden die de enquête invulden, waren gemiddeld hoger geschoold en op zich al geïnteresseerd in goede voedingsgewoontes. Anders hadden ze de vrijblijvende vragenlijst nooit ingevuld.

- “Computerapparatuur en –programmatuur in het Vlaams katholiek secundair onderwijs” is een brochure van het VVKSO over de ICT-situatie op 1 januari 2004. Het was de bedoeling om 593 onderwijsinstellingen te bevragen. Daarvan hebben er 348 geantwoord op het toegestuurde enquêteformulier. In die brochure staat daarover onderstaande tekst. Die wijst op een mogelijk probleem. Welk? Is het verstandig om daar de aandacht op te trekken? Kan je het woord “respondenten” in zijn juiste context plaatsen?

We moeten opmerken dat het beeld dat uit de enquête naar voren komt, ongetwijfeld te rooskleurig is omdat de steekproef niet aselekt is. Immers, scholen die het gevoel hebben mee te zijn met ICT, zullen eerder geneigd zijn de enquête te beantwoorden dan scholen die het gevoel hebben achterop te hinken in ICT. Bij de respondenten zijn de beter-dan-gemiddeld-scholen dus waarschijnlijk sterker vertegenwoordigd.

- In de reeds vermelde brochure van het VVKSO staat onderstaande grafiek. Is het een gepaste grafiek voor het soort veranderlijke dat wordt getoond? Kan je iets zeggen over de globale vorm? Zijn er dingen die beter of duidelijker kunnen? Wat kan je uit de figuur afleiden over het gemiddelde en de mediaan?



- Hieronder staan de samengevatte resultaten van 2 onderzoeken, uitgevoerd in Diest en in Westmalle. De steekproeven waren niet even groot. Kan je voor beide onderzoeken het gemiddelde berekenen? Kan je ook de mediaan bepalen? En kan je tenslotte die twee onderzoeken grafisch met elkaar vergelijken in eenzelfde figuur?

<i>Aantal honden Diest</i>	<i>Frequentie</i>
<i>0</i>	<i>7</i>
<i>1</i>	<i>4</i>
<i>2</i>	<i>3</i>
<i>3</i>	<i>1</i>

<i>Aantal honden Westmalle</i>	<i>Relatieve frequentie</i>
<i>0</i>	<i>0.48</i>
<i>1</i>	<i>0.32</i>
<i>2</i>	<i>0.16</i>
<i>3</i>	<i>0.04</i>

Een statistisch onderzoek naar het schatten van de tijdsduur van 1 minuut

1 Wat wil je weten? Hoe ga je meten?

1.1 De onderzoeksvraag

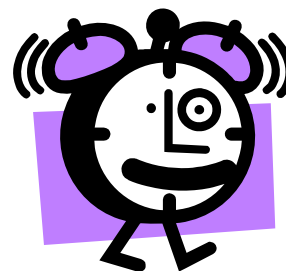
In sporten zoals atletiek, zwemmen, Formule 1, speelt tijdsopname een belangrijke rol. Probeer je eens voor te stellen wat in de cockpit van een Ferrari gebeurt om de rondetijden van Michaël Schumacher tot op een duizendste van een seconde te kunnen opmeten.

De meest primitieve manier van tijdsopname is vermoedelijk gewoonweg ... tellen. Je kent waarschijnlijk het 21 – 22 – 23 trucje om 3 seconden te benaderen, maar de tijdsduur van één minuut schatten is heel wat moeilijker. Of niet? Aan jou de uitdaging om dit te onderzoeken!

Let op! Wat is je populatie?

Je hebt hier opnieuw een probleem. Je moet eerst heel nauwkeurig zeggen wat je bij welke populatie wil onderzoeken. Misschien wil je weten hoe één bepaalde leerling een minuut schat, om te constateren dat zij een kwart van de keren te hoog en drie kwart van de keren te laag schat. Dan bestaat je populatie (in theorie) uit alle resultaten van die leerling, als je die miljoenen keren een minuut zou laten schatten. En als steekproef zal je dan die éne leerling 40 keren laten schatten. Zo krijg je een idee over het schattingsgedrag van die éne leerling.

Maar je kan ook iets anders willen weten. Hoe schatten de leerlingen op je school, als zij één keer de kans krijgen om te schatten? Je populatie bestaat dan uit alle leerlingen van je school. Daaruit kan je een steekproef trekken van 40 leerlingen, die één keer de kans krijgen om een minuut te schatten. We spreken af dat je dit tweede probleem gaat onderzoeken.



- *Wat onderzoek je hier bij welke populatie?*

- *Eigenlijk geef je een kleine opdracht aan elke leerling uit je steekproef. Is het belangrijk dat je vooraf vastlegt hoe die opdracht moet uitgevoerd worden? Wat is de afspraak die je maakt om het onderzoek correct uit te voeren? Kan de plaats of het tijdstip van ondervragen invloed hebben op de kwaliteit van de metingen?*

- Kan je de steekproef die je getrokken hebt vooraleer je aan het derde onderzoek begon gebruiken voor je huidige onderzoeksvraag? Waarom? Wat is je dataset hier?

1.2 De dataset: getallen en context.

Je dataset bevat de geschatte tijdsduur, uitgedrukt tot op de seconde. Deze veranderlijke kan (minstens in theorie) oneindig veel verschillende waarden aannemen die willekeurig dicht tegen elkaar kunnen liggen. Een mogelijke geschatte waarde kan 54.374 seconden zijn. In dit onderzoek is afgesproken dat je de opmetingen afrondt tot op de seconde. Maar het is niet omdat jij met zo'n afronding werkt (en dus 54 seconden opschrijft) dat de echte tijd ook met sprongen verloopt.

Als je te maken hebt met uitkomsten die alle mogelijke getalwaarden kunnen aannemen tussen bepaalde grenzen, dan spreek je over een “**continu numerieke**” veranderlijke. Je weet al dat de naam “**numeriek**” wijst op het feit dat je echt met getallen te maken hebt, en niet met landen of wielersponsors. De naam “**continu**” wijst erop dat (minstens theoretisch) de getallen alle mogelijke waarden kunnen aannemen tussen bepaalde grenzen, zonder enige onderbreking.

De geschatte tijdsduur van één minuut is een voorbeeld van een **continu numerieke veranderlijke**. Een ander voorbeeld is het gewicht van een leerling of haar lengte.



Als je continue gegevens opschrijft, dan moet je altijd ergens afronden. Het lijkt er dan op dat tussen de verschillende mogelijke waarden ook tussenstappen zijn, net zoals bij de discreet numerieke veranderlijke, maar die tussenstappen in jouw getallen zijn een gevolg van afrondingen. Bij de echte waarden zijn er geen vaste tussenstappen, en daarom noemt men zo'n veranderlijke continu.

Dit is een nieuw type veranderlijke, waarvoor een nieuwe werkwijze nodig is bij het opstellen van de frequentietabel en het tekenen van de bijhorende grafieken.

2 Op speurtocht in je dataset

Om een goed zicht te krijgen op al je verzamelde gegevens, zal je de getallen uit de dataset samenvatten in een tabel. Je zal ook grafieken tekenen en kengetallen berekenen. Je GRM komt weer goed van pas.

Neem je GRM en tik al de opgemeten tijden in de lijst [L1]. De volgorde waarin je die waarnemingen invoert heeft geen belang.

L1	L2	L3	1
59	-----	-----	
65			
69			
72			
61			
56			
56			
L1(1)=59			



2.1 Een frequentietabel met klassenindeling

Je beschikt nu over een groot aantal opmetingen van een continu numerieke veranderlijke. In feite is elke geschatte tijdsduur verschillend van elke andere, maar daarvoor had je moeten meten tot op een miljardste van een seconde (of misschien nog preciezer!). Als elke “echte” waarde verschillend is van elke andere, dan komt elke “echte” waarde slechts één keer voor. Een frequentietabel zou dan (theoretisch) al die verschillende “echte” waarden moeten bevatten, allemaal met een frequentie gelijk aan één. Dat is zinloos.

De tijdsmetingen die je hier hebt, worden, zoals alle continue veranderlijken, samengevat in een **frequentietabel met klassenindeling**.

Klasse	Frequentie
[30; 35[1
[35; 40[0
[40; 45[...
...	...

Voor het maken van de klassen kan je als volgt te werk gaan.

- Start met een interval dat groot genoeg is om al je opmetingen te kunnen bevatten. Als je kleinste observatie 32 is en je grootste is 93, dan moet je dus minstens van 32 tot 93 gaan. Meestal neem je eenvoudige “ronde” getallen. Hier zou je bijvoorbeeld kunnen starten bij 30 en eindigen bij 95 of 100.
- Op dit grote interval maak je nu deelintervallen die mooi aan elkaar aansluiten en elkaar niet overlappen. Dat zijn je klassen. De breedte van die klassen mag je zelf kiezen en ze hoeven zelfs niet allemaal even breed te zijn.
- Elke klasse is een “links gesloten – rechts open” interval, zoals bijvoorbeeld [30 ; 35[. De grenzen van een klasse heten **klassengrenzen**. Het midden heet **klassenmidden** en de breedte heet **klassenbreedte**.
- Zorg ervoor dat het overgrote deel van de waarnemingen niet binnen één of twee klassen valt. Als richtlijn neem je tussen de 5 en de 15 klassen, maar deze richtlijn hoeft je niet te strikt te nemen. Als je de frequentietabel gebruikt om een histogram te tekenen (zoals uitgelegd in volgend puntje), dan zal je ervaren dat te veel klassen dikwijls een zeer onrustige figuur geven terwijl te weinig klassen bijna niets meer tonen.

Om een frequentietabel met klassenindeling op te stellen kan je je laten helpen door de GRM.

- Kijk na of al je gegevens in [L1] juist zijn ingevoerd.
- Om een beter overzicht te hebben kan je de gegevens sorteren. Druk [STAT] en kies 2:SortA(. Druk daarna [2nd][L1] en [ENTER].
- De lijst [L1] is nu gesorteerd. In dit onderzoek is de laagst geschatte tijdsduur 32 seconden. De hoogste is 93 seconden.

Bij continu numerieke gegevens is het mogelijk om met behulp van je GRM een frequentietabel met klassenindeling op te stellen. Je maakt hiervoor een kleine omweg door eerst een tekening te maken van je dataset. Op die manier moet je niet meer turven!

- Gebruik **2nd** [STAT PLOT]. Zorg ervoor dat alle Plots op Off staan. Kies 1:, kies On en **ENTER**, dan **1** en **ENTER**, dan Xlist: [L1] en Freq: 1

```

5:HiPlot3
1:Plot1...Off
  L1 L2
2:Plot2...Off
  L1 L2
3:Plot3...Off
  L1 L2
4:PlotsOff

```

```

2001 Plot2 Plot3
Off Off
Type:
Xlist:L1
Freq:1

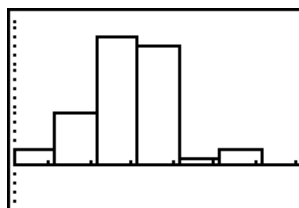
```

- Druk **ZOOM** en kies 9:Zoomstat. Je GRM heeft nu een histogram van je dataset getekend met zelf gekozen klassen

```

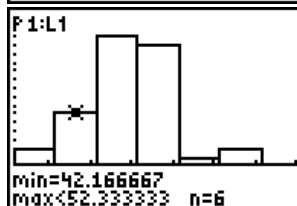
2001 MEMORY
3:Zoom Out
4:ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZoomStat

```



- Druk **TRACE** en gebruik **←** en **→** om de figuur te doorlopen. Je ziet bijvoorbeeld klassengrenzen van 42.16.. tot 52.33...

Druk **WINDOW** en kijk hoe je GRM die klassen maakt. Xscl bepaalt de klassenbreedte, en daar staat 10.166... Zoiets wil jij natuurlijk niet en dus pas je die instellingen aan. Begin bijvoorbeeld met Xmin=30, Xmax=100, en Xscl=10.



```

WINDOW
Xmin=32
Xmax=103.16666...
Xscl=10.166666...
Ymin=-4.51035
Ymax=17.55
Yscl=1
Xres=1

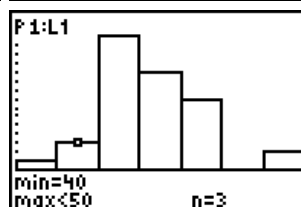
```

Druk dan terug **TRACE**. Doorloop de figuur. Je ziet dat de frequentie van de klasse [40 ; 50[gelijk is aan 3. Dat lees je onderaan af bij n=...

```

WINDOW
Xmin=30
Xmax=100
Xscl=10
Ymin=-4
Ymax=17
Yscl=1
Xres=1

```

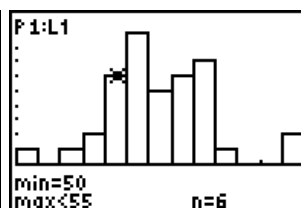


- Om een frequentietabel te maken die begint bij [30; 35[en waarbij elke klasse 5 eenheden breed is druk je **WINDOW** en pas je de instellingen aan zoals hiernaast. Druk dan **TRACE** en doorloop de figuur. Alles wat je nodig hebt om een frequentietabel met klassenindeling te maken kan je nu aflezen.

```

WINDOW
Xmin=30
Xmax=95
Xscl=5
Ymin=-3
Ymax=10
Yscl=1
Xres=1

```



- Stel nu een frequentietabel op voor jouw onderzoek. Je kan je laten leiden door bovenstaand voorbeeld en de klassenbreedte gelijk aan 5 nemen, tenzij dat voor jouw data niet zinvol is.

Klasse	Frequentie

2.2 Het histogram

Een basisfiguur om continu numerieke gegevens te onderzoeken is het **histogram**.



Raadpleeg eerst paragraaf 3.3 in je infoboekje om te leren hoe je een histogram moet tekenen.



Er zijn geen vaste regels voor het aantal klassen en dus kan je veel verschillende histogrammen tekenen voor eenzelfde dataset. Probeer zelf enkele klassenbreedtes uit. Je zal telkens verschillende histogrammen zien. Let daarbij op de **globale** vorm en probeer daaruit kenmerken van de onderzochte dataset te ontdekken.

Basisafpraak voor het tekenen van een histogram.

De OPPERVLAKTE van een rechthoek is recht evenredig met het aantal observaties in de klasse waarop die rechthoek staat

Bemerk dat de oppervlakte niet de fysische oppervlakte van het balkje is dat je op je blad papier hebt getekend. Je moet dus niet je lat bovenhalen en beginnen meten. De oppervlakte die men hier bedoelt is het product van de basis, die je afleest op de x-as, en de hoogte, die je afleest op de y-as. We houden ons daarbij niet bezig met eenheden, we kijken alleen naar het maatgetal van de oppervlakte.

Je GRM tekent altijd histogrammen met gelijke klassenbreedte. Als maatgetal voor de hoogte van de rechthoeken neemt de GRM de frequentie. Dat betekent dat de evenredigheidsfactor k altijd gelijk wordt genomen aan de vaste klassenbreedte. Dit kan je (eventueel als extra oefening) eenvoudig nagaan. Bestudeer daarvoor de oppervlaktes en vergelijk die met de frequenties.



Sommigen denken dat een histogram en een staafdiagram goed op elkaar lijken. Dat is fout, want er zijn fundamentele verschillen. Een histogram hoort bij een continu numerieke veranderlijke waar geen tussenstappen zijn tussen de “mogelijke” uitkomsten. Bij een histogram liggen de rechthoeken dus tegen elkaar. Bij een staafdiagram is er open ruimte tussen de staafjes. Bovendien kijk je bij een staafdiagram naar de hoogte en bij een histogram naar de oppervlakte.

- *Teken een histogram voor de door jou opgestelde frequentietabel.*



2.3 Numerieke kenmerken

2.3.1 Gemiddelde en mediaan



Hoe de testpersonen de tijdsduur van 1 minuut geschat hebben is mooi weergegeven in het histogram. Net zoals bij de discreet numerieke veranderlijke kan je nu ook een aantal kengetallen berekenen.

Als eerste kenmerk wil je bepalen hoeveel seconden een proefpersoon “typisch” heeft geschat. De gebruikelijke kengetallen hiervoor zijn het gemiddelde en de mediaan. Je kan die berekenen met je GRM.

- Al je gegevens staan in [L1]. Activeer **2nd** [LIST], kies MATH en dan 3: mean voor de berekening van het gemiddelde en 4: median voor de berekening van de mediaan. Vul het commando aan met **2nd** [L1]. Bevestig met **ENTER**.

NAMES OPS MATH	mean(L1	61.35
1:min(median(L1	61
2:max(
3:mean(
4:median(
5:sum(
6:Prod(
7:stdDev(

- Noteer nu de centrummaten. Gebruik de juiste symbolen. Vergeet de eenheid niet.*

2.3.2 Standaardafwijking en interkwartielafstand

Om je opmetingen te karakteriseren is het “centrum” maar een eerste stap. Een tweede karakteristiek is de spreiding rond dit centrum. De gebruikelijke kengetallen hiervoor zijn de standaardafwijking en de interkwartielafstand. Zij worden ook **spreidingsmaten** genoemd.

De standaardafwijking

Om de spreiding van de gegevens rond het gemiddelde te berekenen, gebruik je de **standaardafwijking s**. Voor de standaardafwijking bestaat een “te gekke” formule. Die ziet eruit als

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Gelukkig kan je dit kengetal berekenen met de GRM.



Raadpleeg paragraaf 3.2 in je infoboekje voor meer informatie over de standaardafwijking.

Je gegevens staan nog steeds in lijst [L1]. Activeer **2nd** [LIST], kies MATH en dan 7:stDev voor de berekening van de standaardafwijking. Vul aan met **2nd** [L1] (*stDev staat voor standard Deviation*). Voor dit onderzoek is $s = 11.84$ seconden.

NAMES OPS MATH	stdDev(L1	11.83985014
1:min(
2:max(
3:mean(
4:median(
5:sum(
6:Prod(
7:stdDev(

De interkwartielafstand

Een andere maat voor spreiding is de lengte van het gebied waarin de middelste 50% van de geordende opmetingen liggen. Dit gebied loopt van het eerste kwartiel Q_1 tot het derde kwartiel Q_3 . De lengte van dit interval is de interkwartielafstand, genoteerd als IQR (= *InterQuartile Range*). De IQR kan je met je GRM bepalen.



Lees nu eerst paragraaf 4.1 van het infoboekje met extra informatie over de kwartielen.

Ga via [STAT] naar CALC, 1:1-Var Stats en vervolledig met [L1], [ENTER]. Loop met ∇ naar beneden.

Hier vind je naast het minimum en het maximum ook de mediaan en de kwartielen.

```
EDIT 1-Var Stats
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
```

```
1-Var Stats L1
```

```
1-Var Stats
n=40
minX=32
Q1=55
Med=61
Q3=69.5
maxX=93
```

In dit voorbeeld is de $IQR = Q_3 - Q_1 = 69.5 - 55 = 14.5$ seconden. De middelste helft van de geordende opmetingen ligt in het interval [55 ; 69.5].

- Noteer de kwartielen en de spreidingsmaten. Vergeet niet dat deze grootheden een eenheid hebben.

2.4 De boxplot

Een goed zicht op zowel het centrum als de spreiding van je opmetingen, krijg je uit een boxplot. Dit is een grafiek die gebruik maakt van de begrippen *minimum*, *maximum*, *mediaan*, *eerste kwartiel* Q_1 , *derde kwartiel* Q_3 , *IQR* en *uitschieters*.

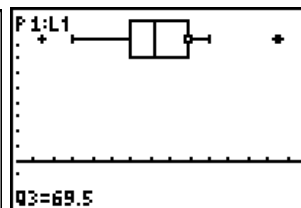


Lees nu eerst paragraaf 4.2 van het infoboekje met extra informatie over de boxplot.

Met de GRM teken je een boxplot als volgt.

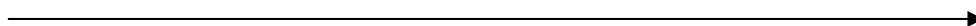
- Druk [2nd] [STAT PLOT] en kijk of alle Plots op Off staan. Activeer dan 1:Plot1 en vervolledig zoals hiernaast. Druk dan op [ZOOM] en 9: ZOOMSTAT
- Probeer ook eens [TRACE].

```
Plot1 Plot2 Plot3
On Off
Type: [L1] [L2] [L3]
Xlist: L1
Freq: 1
Mark: [ ] [ ] [ ]
```



- *Hoe lang zijn de “staarten”? Wat betekenen de “rechthoekjes”?*

- *Teken nu de boxplot. Vergeet de x-as niet te voorzien van de juiste getallen en de juiste eenheid.*



2.5 Histogram en boxplot interpreteren

Zodra je een histogram of boxplot hebt getekend probeer je daar zoveel mogelijk informatie uit af te lezen.



De manier waarop je een histogram tekent, heb je zelf in de hand. Als je een andere keuze maakt voor het aantal klassen of voor de klassenbreedte, dan krijg je een andere figuur. Bij een boxplot is dat niet zo: je kan voor een bepaalde dataset maar één boxplot tekenen.

Een combinatie van beide figuren is dikwijls interessant om goede conclusies te trekken.

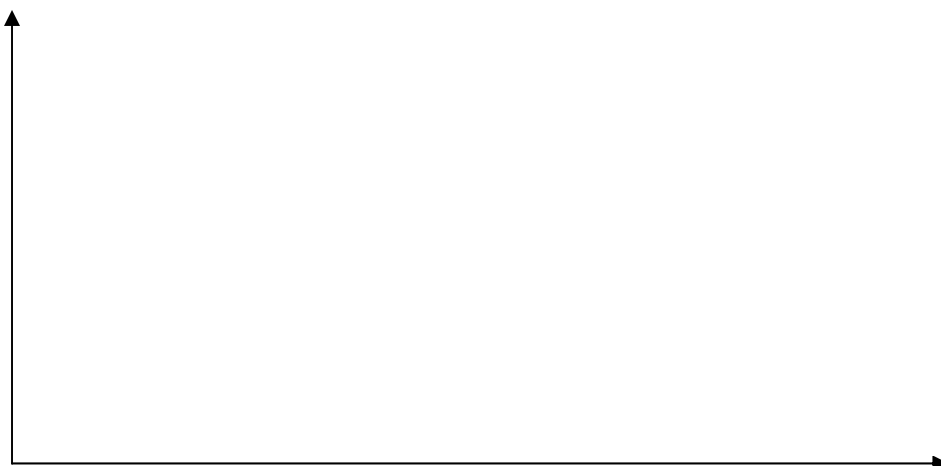
De combinatie van het histogram en de boxplot helpt je om je dataset te interpreteren.

- *Is het histogram symmetrisch of scheef? Is die scheefheid heel sterk of maar een klein beetje? Kan je die informatie ook uit de boxplot halen? Naar wat kijk je dan?*

- *Zijn er uitschieters in je dataset? Was dat te verwachten?*
- *De mediaan is het midden van je geordende dataset. Je kan die duidelijk zien op de boxplot. Teken nu eens een histogram waarbij de mediaan een klassengrens is. Neem als klassenbreedte bijvoorbeeld 10. Zie je hier een duidelijke scheefheid? Gebruik je GRM.*



- *Probeer ook eens met klassenbreedte 5. Wat bemerk je in vergelijking met je eerste histogram met klassenbreedte 5 ? Gebruik je GRM.*



- *Had je op basis van die histogrammen kunnen zeggen welk kengetal, mediaan of gemiddelde, het grootst zou zijn? Waarom?*

3 Wat heb je gevonden? Hoever kan je gaan in je conclusie?

3.1 De variabiliteit van steekproefresultaten

Je bent er nu al mee vertrouwd. Een steekproef levert toevallige resultaten op, en bij een andere steekproef krijg je andere resultaten. Indien je de populatie van alle leerlingen van je school in identieke omstandigheden de tijdsduur van één minuut zou laten schatten, dan zou het gemiddelde van al die schattingen niet exact samenvallen met het gemiddelde dat jij in je steekproef hebt gevonden. Dat is niet erg. De statistiek is er juist om je te helpen om goede uitspraken over de populatie te doen. Als er tenminste geen andere problemen opduiken... .

In dit onderzoek heb je elke leerling aan een kleine test onderworpen. De manier waarop die test moet worden afgenomen heb je vooraf heel precies vastgelegd, en je hebt je aan die procedure strikt gehouden bij elke leerling die je hebt getest. Maar er is nog iets anders dat voor fouten kan zorgen. Je werkt met meetapparatuur, en is die wel goed geijkt? Als je een chronometer gebruikt die start bij 2 in plaats van bij 0, dan heb je in al je opmetingen een systematische fout van 2 seconden. Je krijgt dan een vertekend beeld van de werkelijkheid.

Als al je opmetingen op een systematische manier te klein (of te groot) zijn dan heb je **vertekening**. Vertekening in metingen kan je met statistiek niet opsporen. Je moet vooraf controleren of je apparatuur wel juist geijkt is. Doe dit vooraleer je aan je onderzoek begint!

- *Soms kan je op een spitsvondige manier vertekening in opmetingen neutraliseren. Als je een weegschaal moet gebruiken waarvan je vermoedt dat zij systematisch een te laag gewicht aangeeft, hoe zou jij dan een boekentas daarop wegen, als je met die weegschaal vooraf niets mag doen?*
- *Je moet de lengte van 20 planken meten en je doet dat met een rolmeter. Op welke manier zou hier vertekening kunnen optreden?*

3.2 Enkelvoudig aselekt, en nog veel meer

Steekproeven, waarbij je het toeval op een gecontroleerde manier zijn rol laat spelen, heten “toevalsgestuurde” steekproeven. Zo zijn er verschillende soorten.

Een eerste is de enkelvoudige aselekte steekproef (afgekort als EAS). Dat is de basis, en die ken je al. Door lukraak een groepje van 40 leerlingen te trekken uit de populatie van 512 leerlingen, heeft elk groepje van 40 leerlingen dezelfde kans om jouw steekproef te zijn.

Er zijn ook andere manieren om aan 40 leerlingen te komen. We bekijken er ééntje van. Start met een lijst van alle klassen van je school. Nummer die klassen en trek daaruit een EAS van 8 klassen. Voor die 8 klassen vraag je de namen van de leerlingen. Per klas trek je een EAS van 5 leerlingen. Zo heb je ook 40 leerlingen in je steekproef. Deze manier van steekproeftrekken gaat in

stapjes, en wordt daarom “getrapt” genoemd. Je kan dit systeem natuurlijk uitbreiden. Als je 400 leerlingen uit Vlaamse scholen wil, dan kan je bijvoorbeeld eerst lukraak 20 scholen trekken, dan in elke school lukraak 5 klassen, en dan in elke klas lukraak 4 leerlingen. Dit is opnieuw een voorbeeld van een “getrapte steekproef”.

- *Als je voor je huidig onderzoek zo weinig mogelijk klassen wil “storen”, werk je dan met een EAS of met een getrapte steekproef? Waarom?*
- *Jij neemt een EAS van 40 leerlingen in je school. Je leerkracht gebruikt een getrapte steekproef (eerst 8 klassen, dan 5 leerlingen per klas) om aan een steekproef van 40 leerlingen te komen. Zijn dit gelijkwaardige methoden (kan je leerkracht alle groepjes van 40 uitkomen die jij kan uitkomen en kan jij alle groepjes van 40 uitkomen die je leerkracht kan uitkomen)?*

3.3 Een uitspraak over de populatie

Wat je in je “exploratief” onderzoek hebt gevonden is van toepassing op de dataset die jij hebt onderzocht, en dus op die 40 leerlingen. Als je op een goede manier een steekproef trekt, als je je strikt houdt aan de procedure om de leerlingen te testen, en als je meetapparatuur goed geijkt is, dan kan je met statistiek verantwoorde uitspraken doen over hoe heel de school een minuut zou schatten. Voor deze steekproef was het gemiddelde 61.4 seconden en de mediaan was 61 seconden. Dat ligt niet ver uit elkaar, en je zou kunnen vermoeden dat het gemiddelde en de mediaan van de hele populatie ook wel in de buurt van 61 seconden liggen. Waarschijnlijk is dat nog waar ook.

4 Kernachtige samenvatting van dit onderzoek

Je samenvatting bestaat uit twee delen

- *De antwoorden op de contextvragen (de www-vragen)*
- *De besluiten over het uitgevoerde onderzoek*

Betrek ook de kengetallen in je besluit: vermeld hun getalwaarde en ga na in hoever ze een zinvolle karakteristiek zijn voor dit onderzoek. Vergeet ook nooit om goede grafieken te tekenen en die te interpreteren.

- *Formuleer nu de antwoorden op de contextvragen (de www-vragen)*

- *Formuleer nu de besluiten over het uitgevoerde onderzoek.*

5 Zelfevaluatie

In dit onderzoek heb je geleerd over:

- getrapte steekproeven
- vertekening bij opmetingen
- de continu numerieke veranderlijke
- de frequentietabel met klassenindeling
- het histogram en de boxplot
- de standaardafwijking en de interkwartielafstand
- de interpretatie van kengetallen in combinatie met grafieken

Je bent nu in staat om volgende opdrachten uit te voeren:

- *Wanneer is een numerieke veranderlijke continu? Zeg dat in je eigen woorden, en geef enkele voorbeelden.*

- *Schrijf je een continu numerieke veranderlijke altijd op met kommagetallen? Motiveer je antwoord.*

- *Welke eigenschap probeert de standaardafwijking te beschrijven? Zeg in woorden hoe je de standaardafwijking berekent. Kan je daaruit afleiden of de standaardafwijking gevoelig is voor uitschieters? Kan je daarvan een eenvoudig voorbeeld geven (je mag je GRM gebruiken)?*

- *Welke eigenschap probeert de interkwartielafstand te beschrijven? Zeg in woorden wat kwartielen zijn en hoe je de interkwartielafstand berekent. Kan je daaruit afleiden of de interkwartielafstand gevoelig is voor uitschieters? Kan je daarvan een eenvoudig voorbeeld geven (je mag je GRM gebruiken)?*

- *De leesbaarheid van een tekst hangt ondermeer af van de lengte van de zinnen. Korte zinnen lezen gemakkelijker dan lange zinnen. Als je aanneemt dat zowat elke nieuwe zin met een hoofdletter begint, en dat er verder niet te veel afkortingen in hoofdletters voorkomen, dan is de verhouding van het aantal hoofdletters ten opzichte van het totale aantal letters een goede maat voor de lengte van de zinnen. Het is nu aan jou om voor deze werktekst (die ongeveer 40 bladzijden telt) een goede schatting te maken van de proportie hoofdletters. Hoe ga je dat doen? Gebruik jij een EAS? Wat zou je dan moeten doen? Gebruik jij een getrapte steekproef? Hoe zou jij dat dan doen? Zijn er verschillende mogelijkheden?*

- *Stel de volgende frequentietabel grafisch voor met behulp van een histogram. Denk daarbij goed aan de basisafspraken voor het tekenen van histogrammen. Je mag gebruik maken van het programma HISDICH samen met zijn handleiding (te downloaden vanaf www.uhasselt.be/lesmateriaal-statistiek).*

Het programma HISDICH maakt een speciale keuze voor de evenredigheidsfactor. Welke? Zoek dit uit door de oppervlaktes te vergelijken met de frequenties.

klassen	frequentie
[20; 40[8
[40; 50[8
[50; 60[24
[60; 70[21
[70; 80[16
[80; 90[14
[90; 120[9