

## Hoofdstuk 2

### Inleidende begrippen

#### 1. Inleiding

Je doet criminologisch onderzoek om iets over de werkelijkheid te weten te komen. We kunnen geïnteresseerd zijn in daders, slachtoffers, de bevolking en diens attitudes ten aanzien van misdaad en straf, en de actoren van het strafrechtelijk apparaat: de politie, de openbare aanklager, de rechters, het gevangeniswezen,... Op basis van jouw onderzoek kun je dan uitspraken doen over een element van de werkelijkheid. Wie statistisch onderzoek doet, gelooft tot op zekere hoogte dat er een werkelijkheid bestaat die wij tot op zekere hoogte kunnen meten en proberen te begrijpen. Daarbij moet duidelijk worden over wie of wat je op basis van het criminologisch onderzoek een uitspraak doet. Dat zijn de **objecten** of **onderzoekseenheden**, de personen of zaken over wie je iets zegt. Als je op basis van een onderzoek bijvoorbeeld de conclusie trekt dat de gemiddelde leeftijd van de eerstejaarsstudenten van een universiteit 19,7 jaar is, dan zeg je op basis van dit onderzoek iets over eerstejaarsstudenten. Dat zijn de onderzoekseenheden. Van deze studenten beschrijf je een kenmerk, namelijk de leeftijd van de studenten. In het onderzoek kun je meer kenmerken van de studenten hebben verzameld, zoals geslacht, vooropleiding en studiekeuze. Deze **kenmerken** (leeftijd, geslacht, vooropleiding en studiekeuze) zijn de **variabelen** in het genoemde onderzoek. Een criminoloog zou zich echter ook de vraag kunnen stellen hoeveel pintjes de student drinkt tijdens een week, of hoeveel keer een student medestudenten besteelt of spiekt tijdens het examen. Wie dit onderzoekt, komt tot de vaststelling dat ook criminologiestudenten wel eens uit de bocht vliegen. Logisch, studenten zijn ook maar mensen, en niks menselijks kan hun dus vreemd zijn. Toch wordt het overtreden van bepaalde regels niet getolereerd. Daarom is het interessant te kijken hoe vaak bepaalde vormen van regelovertreding voorkomen, en welke de achtergrondkenmerken zijn van de regelovertreders.

Onderzoekseenheden hoeven niet altijd personen te zijn. Je kunt op basis van een onderzoek ook uitspraken doen over de lengte van misdaadartikelen in dagbladen. In dat geval zijn de artikelen over criminaliteit de onderzoekseenheden, de objecten waarover je een uitspraak doet. De lengte is hier een kenmerk van de onderzochte artikelen en is daarom een variabele in het onderzoek. Lengte kan worden uitgedrukt in het aantal woorden of tekens. Met behulp van statistiek kun je precieze uitspraken doen over de kenmerken van onderzoekseenheden, zoals 'de gemiddelde leeftijd van eerstejaarsstudenten is 19,7 jaar', 'de proportie van

studenten die dronken achter het stuur kruipt is één op tien (0.1) en ‘de meeste misdaadartikelen zijn niet langer dan 600 woorden.’

Een ander voorbeeld: ‘Televisiekijkers met een hoge opleiding kijken vaker naar het nieuws dan televisiekijkers met een lage opleiding.’ In deze uitspraak wordt iets gezegd over televisiekijkers. In het onderzoek zijn televisiekijkers de onderzoekseenheden. In het voorbeeld zijn de kenmerken van die televisiekijkers: opleiding en de frequentie waarmee naar het nieuws wordt gekeken. ‘Opleiding’ en ‘frequentie nieuws kijken’ zijn de variabelen in het onderzoek.

Gegevens verzamelen die voor statistische analyse bruikbaar zijn, is een grote opgave. De individuen waarover je informatie wilt verzamelen, dienen op voorhand duidelijk gedefinieerd te zijn. Willen we informatie over een volledige bevolking (bijvoorbeeld alle Belgen) of nemen we genoegen met een deelgroep, bijvoorbeeld alle Belgen tussen 15 en 90 jaar? De onderzoeker bepaalt dit. Maar we mogen achteraf niet vergeten hoe we de bevolking hebben gedefinieerd. Anders bestaat de kans op overgeneralisatie! Deze fout wordt in al het enthousiasme over de onderzoeksresultaten nog veel te vaak gemaakt. De verzameling van individuen waarover we een uitspraak willen doen is de **onderzoekspopulatie**. Meestal is het onmogelijk om alle individuen uit onze onderzoekspopulatie te bevragen. In dit geval moeten we het met minder doen. Het zou onmogelijk zijn om bijvoorbeeld alle Belgen tussen 15 en 90 jaar te bevragen over de mate waarin zij zich onveilig voelen. We nemen daarom een staal uit de onderzoekspopulatie. Dit staal noemen we een **steekproef**. De personen die we uiteindelijk bevroegd hebben, zijn de **respondenten** en zijn dus een deelverzameling van de onderzoekspopulatie. Er bestaan verschillende vormen van steekproeven. De diverse soorten steekproeven die bestaan, worden in andere handboeken behandeld. Het volstaat hier te zeggen dat we uitspraken kunnen doen binnen een bepaalde marge als onze steekproef **toevalsgewijs (Engels: at random)** is samengesteld. Belangrijk is ook het **representatief** karakter van de steekproef. Met representativiteit wordt bedoeld dat een kenmerk in de steekproef evenveel voorkomt als in de onderzoekspopulatie. Als een steekproef niet representatief is, is een bepaalde groep oververtegenwoordigd en een andere groep ondervertegenwoordigd. Zo zijn spijbelaars vaker afwezig dan niet-spijbelaars in enquêtes die op school worden afgenomen.

## 2. Beschrijven, schatten en veralgemenen als statistische bedrijvigheid

De statistiek die in deze syllabus aan bod komt, is een wetenschapstak en volgt bijgevolg de hiervoor genoemde criteria om tot wetenschappelijke inzichten te komen. Statistiek kunnen we algemeen omschrijven als: ‘het geheel van regels en procedures om gemeten kenmerken te verwerken’. We onderscheiden twee soorten statistiek:

- **Beschrijvende statistiek:** of het op een overzichtelijke en samenvattende of synthetische wijze weergeven van kenmerken die voorkomen in een onderzoekspopulatie of in een steekproef, wanneer we niet alle eenheden kunnen bevragen.
- **Inductieve of inferentiële statistiek:** het veralgemenen van de gegevens verzameld voor een steekproef naar de onderzoekspopulatie waaruit ze getrokken werden. Kunnen we bijvoorbeeld zeggen dat 5% van de jonge adolescenten uit de eerste graad van het secundair onderwijs betrokken is bij geweld in groepsverband als we een steekproef hebben genomen? Of moeten we de onzekerheid die voortvloeit uit het feit dat we slechts één steekproef hebben genomen meenemen om een marge te bepalen, waarbinnen we met een zekere graad van zekerheid kunnen zeggen dat dit kenmerk in de onderzoekspopulatie zal voorkomen. Dit komt uitvoerig aan bod in de verdere hoofdstukken over inductieve statistiek. We maken er in de wereld van het sociaal-wetenschappelijk onderzoek heel dikwijls gebruik van.

De **beschrijvende statistiek** gaat in op haar rol als datareductie-techniek door vooral stil te staan bij de vraag hoe grote hoeveelheden van gegevens zo overzichtelijk mogelijk kunnen gepresenteerd worden, zonder dat er (veel) informatie verloren gaat. Hierbij is het belangrijk zich te realiseren dat men onderhevig is aan twee beperkingen:

- *Ten eerste, beschrijvende statistiek laat toe gegevens te beschrijven voor de groep personen of andere eenheden die onderzocht werden. De resultaten kunnen echter niet veralgemeend worden naar andere personen of eenheden, andere tijden. Een beschrijvend onderzoek biedt een momentopname die uitsluitend geldig is voor de onderzochte groep.*
- *Ten tweede, de statistiek laat niet toe om causaliteit vast te stellen. Causaliteit is complex. Omgekeerd, als een causale relatie bestaat, verwachten we dat we een statistisch verband vinden. Alleen is het verband onvoldoende als bewijs.*

We kunnen wel aan de hand van beschrijvende statistiek mogelijke ideeën opdoen over de samenhang van kenmerken. We kunnen veronderstellingen hebben over de wijze waarop criminologische kenmerken samenhangen, zoals leeftijd en criminaliteit, we kunnen deze veronderstellingen **-hypothesen-** opschrijven en kunnen onze wetenschappelijk geformuleerde veronderstellingen dan via de inductieve statistiek toetsen, dit wil zeggen het waarheidsgehalte ervan nagaan. Bijvoorbeeld, we veronderstellen dat de betrokkenheid van burgers bij bepaalde vormen van criminaliteit heel laag is in de kindertijd, toeneemt tijdens de puberteit en afneemt in de volwassenheid. We kunnen gegevens verzamelen over de leeftijd van de respondenten en de criminaliteit die onze respondenten gepleegd hebben in een bepaalde periode en we kunnen op basis van een statistische analyse van de antwoorden nagaan of onze veronderstelling over het verband tussen leeftijd en criminaliteit opgaat in de steekproef.

De beschrijvende statistiek stelt zich tot doel een grote massa aan gegevens op een interessante wijze te beschrijven en samen te vatten of te synthetiseren. Gegevens worden in categorieën teruggebracht, gemeten, samengenomen, vergeleken en tegenover elkaar afgezet. Sommigen vinden de beschrijvende statistiek slechts reductionistisch: dit wilt zeggen dat men rijke informatie tot naakte cijfers herleidt. Dat is waar, maar dat heeft zijn redenen: de beschrijvende statistiek is maar het begin van een reeks van vragen die de criminoloog zich stelt. Waarom is het zo dat de gemiddelde leeftijd waarop jongeren zich aansluiten bij straatgroepjes veertien jaar is? Wat maakt jongeren in deze leeftijdsgroep vatbaar voor het nemen van dergelijke beslissingen? Bij deze procedures of de bewerking van de gegevens speelt theorie een grote rol. *Statistiek is een tool (een gereedschapskist) die ons helpt de wereld te organiseren en te begrijpen.* Maar dat kan alleen als men nauwkeurig te werk gaat. Welke categorieën gehanteerd worden, wat met elkaar vergeleken wordt, wat en hoe men iets meet, wordt beïnvloed door de wetenschappelijke theorie die men hanteert. In dit handboek gaan we uit van een wetenschappelijke methode voor het verzamelen van gegevens.

**De inferentiële statistiek** tracht bevindingen te veralgemenen naar de onderzoekspopulatie. Criminologen kunnen bepaalde veronderstellingen hebben, bijvoorbeeld over het voorkomen van een kenmerk in de onderzoekspopulatie of over de samenhang tussen een aantal kenmerken. Dergelijke veronderstellingen kunnen we formuleren onder wat we **toetsbare stellingen** noemen. Stellingen kunnen evenwel slechts worden geformuleerd wanneer de onderzoeker de vakliteratuur heeft geëxploreerd, en uit deze exploratie op het spoor is

gekomen van theorievorming rond deze vraag of van eerder verricht onderzoek. Deze stellingen worden **hypothesen** genoemd. *Hypothesen zijn dus specifieke stellingen betreffende de (causale) relatie tussen twee of meer concepten, die afgeleid zijn uit de theorie.* De hypothese omvat nu een **onderzoekshypothese**, een **nulhypothese** en een **alternatieve hypothese**. De onderzoekshypothese formuleert de betrokken stelling in een positieve zin, dit wil zeggen in de richting van de verwachtingen van de onderzoeker. Wie onderzoek wilt doen naar de ruimtelijke spreiding van criminaliteit en zich hierbij laat inspireren door de sociale desorganisatietheorie, zal uit deze theorie toetsbare stellingen dienen te distilleren. We geven een illustratie. Economische achterstelling in buurten leidt tot een toename van geregistreerde criminaliteit in buurten omdat de sociale controle in buurten daalt. De **nulhypothese** is de omkering van de onderzoekshypothese, of de hypothese dat er geen verband bestaat tussen economische achterstelling, sociale controle en criminaliteit in buurten. Indien we vaststellen dat de nulhypothese mag verworpen worden, concluderen we (voorlopig) dat onze onderzoekshypothese, nu alternatieve hypothese genoemd, opgaat.<sup>1</sup>

### **3. Statistiek en de beantwoording van beschrijvende en verklarende onderzoeksvragen**

Criminologische problemen roepen vragen op bij onderzoekers. Om deze specifieke vragen te kunnen beantwoorden, dient een probleemstelling (centrale onderzoeksvraag of vraagstelling) te worden geformuleerd. Om een onderzoek succesvol uit te voeren, is het noodzakelijk die probleemstelling zo nauwkeurig mogelijk te formuleren. Een goed geformuleerde probleemstelling geeft immers richting aan het onderzoek. Deze probleemstelling is de basis van het onderzoek en de wijze waarop deze geformuleerd is, bepaalt de resultaten van het onderzoek.

Met het oog op statistische toepassing kunnen onderzoeksvragen **beschrijvend**, **verkennd** of **verklarend** zijn. Een beschrijvende onderzoeksvraag betreft de kwantitatieve beschrijving van een fenomeen onder studie. Veel criminologisch onderzoek is beschrijvend van aard. Dat

---

<sup>1</sup> Het is nu net via de nulhypothese dat de hypothese zal worden getoetst. Een hypothese kan immers nooit definitief bewezen of bevestigd worden; steeds kan een andere onderzoeker later aantonen dat de hypothese die iemand dacht bewezen te hebben, toch niet opgaat. De toetsing van een hypothese wordt geëffectueerd door een poging tot weerlegging (falsifiëring). Daarom loopt de toetsing van een onderzoekshypothese onrechtstreeks via het onderzoek van de nulhypothese: wanneer de onderzoeker kan aantonen dat de nulhypothese (het omgekeerde van de onderzoekshypothese) niet opgaat, toont hij indirect aan dat de onderzoekshypothese (alternatieve hypothese genoemd) (wellicht) wel opgaat. Dit komt verderop in deze cursus uitvoerig aan bod.

meer jongens dan meisjes delicten plegen (het ‘genderratio’ probleem), en dat vrouwen minder vaak recidive plegen, is reeds talloze malen aangetoond en beschreven (beschrijvend onderzoek), doch nauwelijks verklaard. Dat zoveel criminologisch onderzoek beschrijvend en explorerend blijft, heeft te maken met het feit dat verklarend onderzoek veel ingewikkelder is dan beschrijvend onderzoek. **Verkennend onderzoek** gaat verder dan alleen maar beschrijven. Verkennende onderzoeksvragen hebben betrekking op het aftasten van (nieuwe) veronderstellingen. Men zou kunnen zeggen dat de verkennende onderzoeksvraag zich ergens situeert tussen de beschrijvende en de verklarende. **De verklarende onderzoeksvraag** is erop gericht geobserveerde verschillen of veranderingen in uitkomstvariabelen te verklaren vanuit kenmerken die eigen zijn aan een theoretisch paradigma. De vraag waarom mensen met verschillende frequenties strafbare feiten plegen, is dus een verklarende onderzoeksvraag. Verklarend onderzoek wil een statistische verklaring bieden voor de geobserveerde verschillen tussen eenheden op basis van één of meerdere kenmerken van die eenheden. Het doel van criminologische studies is uiteindelijk het verklaren, maar de stap naar verklaringen is niet altijd mogelijk, zeker niet wanneer rond een verschijnsel erg weinig bekend is. Een eerste stap in het onderzoek is dan vaak eerder beschrijven en verkennen, en dus basisinformatie verzamelen, ten einde later onderzoek naar verklaringen mogelijk te maken. Waar verklarend onderzoek dus op zoek gaat naar het ‘verklaren’ (‘oorzakelijke’ verbanden), geeft beschrijvend onderzoek een opsomming van het voorkomen van bepaalde verschijnselen en brengt het dit voorkomen mogelijk in verband met bepaalde andere verschijnselen, zonder een verklaring te geven voor het fenomeen. Gedegen wetenschappers leiden hun onderzoeksvragen dus (mee) af uit hun grondige kennis van de theorie. Beginnende wetenschappers zoals studenten die een masterproef aanvatten, hebben die gedegen kennis meestal niet. Zij hebben vaak slechts een algemeen idee van een probleemstelling, i.e. de contouren van een probleem, en zijn vaak minder op de hoogte van een criminologische deelsterrein waarop ze een onderzoek zouden willen uitvoeren. Zo zijn veel studenten geïnteresseerd in herstelgerichte justitie, maar zij weten vaak niet erg goed welke specifieke onderzoeksvragen ze beantwoord willen zien. Hun voorstellen zijn dan ook meestal erg breed en dienen grondig ingesnoerd te worden. Een verdiepende literatuurstudie is dan ook onmisbaar. Nogmaals, de statistiek alleen is slechts het vehikel. De inhoud wordt bedacht door de criminoloog.

Hieronder volgen een reeks van voorbeelden van criminologische onderzoeksvragen die beantwoord worden aan de hand van de beschrijvende, verkennende en verklarende statistiek:

- Hoeveel keer in het afgelopen jaar zijn leerlingen uit Gentse basisscholen het slachtoffer geworden van afpersing?
- Zijn er individuele verschillen tussen Vlamingen voor wat betreft de kenmerken tevredenheid tegenover de politie, opvattingen over rechtvaardige behandeling door de politie en vertrouwen in de politie? Zo ja, hoe groot zijn deze verschillen? Om te weten hoe groot deze verschillen zijn maken we gebruik van statistische parameters. Voor elk kenmerk bekomen we een resultaat. Als we de resultaten naast elkaar leggen, komen de verschillen aan de oppervlakte. Als we de verschillen hebben vastgesteld, kunnen we de onderzoeksvragen beantwoorden.
- Hebben hoger opgeleiden minder vertrouwen in politie dan lager opgeleiden?
- Is er samenhang tussen het wantrouwen tegenover politie en het geloof in samenzweringstheorieën?
- Zijn er geografische verschillen in autodiefstal en woninginbraak waar te nemen in Belgische steden en gemeenten? Voor welk fenomeen zijn de verschillen het grootst?
- Kunnen de geobserveerde verschillen tussen buurten in geregistreerde criminaliteit verklaard worden aan de hand van de densiteit en de bevolkingssamenstelling van de buurt?

#### 4. Statistische eenheden

Statistische eenheden zijn de **onderzoekseenheden** waar men een uitspraak over wilt doen. Deze eenheden zijn soms inwoners van een bepaald gebied, studenten of leerlingen, verdachten of gedetineerden. Het gaat lang niet altijd over individuele daders of slachtoffers of de bevolking. Soms gaat criminologisch onderzoek over huishoudens, over scholen, over buurten of gemeenten, of nog hogere aggregatieniveaus zoals arrondissementen of landen in internationaal comparatief onderzoek. In strafrechtelijk onderzoek zijn de eenheden soms dossiers of bedrijven. Deze eenheden kan men beschrijven aan de hand van een aantal kenmerken waarin de onderzoeker geïnteresseerd is. Deze kenmerken noemen we variabelen en zijn bijvoorbeeld gebeurtenissen uit het leven van de respondenten uit het onderzoek (bijvoorbeeld het aantal maal dat men slachtoffer geworden is van een misdrijf). In het geval de eenheden hogere aggregaten zijn, zoals steden, betreffen de kenmerken bijvoorbeeld de “criminaliteitsgraad” (het aantal criminele feiten per duizend inwoners) of de “jeugddelinquentiegraad” (het aandeel 12-17 jarigen die gedurende een bepaalde periode door

de politie werd verdacht van een als misdrijf omschreven feit). **Variabelen** zijn met andere woorden **de kenmerken van statistische eenheden die variëren** en die verschillende scores hebben op een bepaald kenmerk.

De eenheden waarover uitspraken gedaan worden, dienen bij voorkeur te verschillen op criminologisch relevante kenmerken. Dit wil zeggen dat ze **variabiliteit** of **spreiding** dienen te vertonen: er is bijvoorbeeld spreiding in crimineel gedrag wanneer de verschillende onderzoekseenheden met een verschillende frequentie crimineel gedrag vertonen. De criminoloog wil dan weten met welke frequentie de onderzoekseenheden crimineel gedrag vertonen, hoe sterk de eenheden wel verschillen in het plegen van delicten, welke de grote tendensen zijn in het plegen van delicten. Als een kenmerk niet varieert is er sprake van een **constante**. Iets wat niet varieert, kan niet gebruikt worden in statistisch onderzoek. Als alle eenheden dezelfde waarde hebben, is het wel heel nutteloos om zich de vraag te stellen of een constante samenhangt met een variabel kenmerk. We komen daar later nog op terug. Onthoud hier alvast de definitie van een constante: een kenmerk dat niet varieert, of een kenmerk waarop alle eenheden dezelfde waarde hebben.

Waarom heeft onderzoek weinig zin wanneer de onderscheiden per kenmerk niet over de eenheden zijn gespreid?

- Indien bijvoorbeeld alle mensen in een steekproef huiseigenaar zouden zijn, kunnen we de vraag niet beantwoorden of eigenaars beter te betrekken zijn bij een buurt- en informatienetwerk dan huurders.
- Voldoende spreiding betekent dat er per kenmerk tenminste twee verschillende waarden zijn én dat de eenheden verspreid zijn over deze categorieën of waarden van die kenmerken. In het gehanteerde voorbeeld betekent dit dat er naast eigenaars ook een aantal huurders zijn.

Bij de planning van onderzoek moet worden toegezien op voldoende spreiding binnen de kenmerken. We weten uit wetenschappelijk onderzoek dat de spreiding van bijvoorbeeld studieprestaties, verdeling van inkomens, politieke opvattingen meestal voldoende groot is. Maar dit kan problematisch zijn in een aantal criminologiethema's: de spreiding van de slachtoffers, de daders, gevangenen, enz. concentreren zich vooral in één bepaald onderscheid. Zo weten we dat 6% van de Belgische huishoudens in het afgelopen jaar slachtoffer werd van een woninginbraak. De spreiding in dit kenmerk is beperkt en levert



steeds moeilijkheden op bij verdere studie. Indien de spreiding onvoldoende is over de verschillende categorieën van variabelen, kan men soms beslissen het kenmerk uit het onderzoek te schrappen, het aantal eenheden in de steekproef op te drijven (= meer inspanningen doen om meer huishoudens die slachtoffer werden van woninginbraak te betrekken) of de probleemstelling te herzien.

## 5. Univariate, bivariate en multivariate beschrijvende analyse

Het soort analyse die je uitvoert, is mede afhankelijk van de *hoeveelheid variabelen* die je bij je analyse wilt betrekken. Een **univariate analyse** is een analyse van *één kenmerk* dat varieert. Bij de univariate analyse komt het er op neer de verdeling van een kenmerk accuraat te beschrijven. In het volgende hoofdstuk bespreken we in hoofdzaak de centrummaten (zoals daar zijn: het rekenkundig gemiddelde, de mediaan, de modus) en spreidingsmaten (zoals daar zijn: de spreidingsmaat D, de variatie, variantie en standaardafwijking of standaarddeviatie).

Wanneer *twee variabelen* (die we doorgaans aanduiden met “x” en “y”) met elkaar in verband gebracht worden, spreken we van een **bivariate analyse**. We zijn bij de bivariate analyse geïnteresseerd in de samenhang tussen twee kenmerken. Beschouwen we twee kenmerken van buurten als onderzoekseenheden: het criminaliteitsniveau en het werkloosheidsniveau. Een pertinente vraag in de criminologie is: gaat het criminaliteitsniveau van een buurt samen met het niveau van sociale ongelijkheid? Of is de samenhang tussen de criminaliteitsgraad en de bevolkingsdichtheid groter? Is er een verband tussen de moraliteitsgraad en de graad voor geweld? Dit zijn allemaal voorbeelden van een onderzoeksvraag die wordt beantwoord aan de hand van een bivariate analyse. Een andere voorbeeld is: worden jongeren frequenter slachtoffer van geweld dan ouderen? Wanneer je *meer dan twee variabelen* gebruikt, voer je een **multivariate analyse** uit. De multivariate analyse maakt beperkt deel uit van dit handboek. Als student is het enorm belangrijk te weten dat het niet volstaat om de relatie tussen twee kenmerken te bestuderen. Dit kan in bepaalde gevallen tot een foutieve conclusie leiden. Op het einde van dit handboek wordt dit intuïtief duidelijk gemaakt en worden de meest basale multivariate analysetechnieken beschreven. **Onderzoeksvragen in de kwantitatieve criminologie worden behandeld aan de hand van univariate, bivariate en multivariate analyses.**

## 6. Meetniveaus van variabelen

De wijze waarop je een variërend kenmerk meet, bepaalt het meetniveau van de variabele. Het meetniveau van de variabele bepaalt onder andere welke statistische analysetechnieken mogelijk zijn. Hoe hoger een meetniveau, hoe meer mogelijkheden de onderzoeker ter zijner beschikking heeft. Anders gezegd: er zijn meer opties voor de beschrijving van kenmerken gemeten op hogere meetniveaus dan voor de bestudering van kenmerken op lagere niveaus. Er zijn vier meetniveaus: het **nominaal meetniveau**, het **ordinaal meetniveau**, het **interval meetniveau** en het **ratio meetniveau**. Het nominale en ordinale meetniveau vormen samen het *categorische meetniveau*, terwijl het interval en ratio meetniveau samen het *metrische meetniveau* vormen.

### *Het nominale meetniveau en het ordinale meetniveau*

Het **nominale meetniveau** brengt de kenmerken van onderzoekseenheden onder in *elkaar uitsluitende categorieën*. Hier heeft de waarde die een variabele kan aannemen alleen de betekenis van een naam, een categorie. Men spreekt daarom van ‘*categorische gegevens*’. Categorieën maken is eigenlijk hetzelfde als onderzoekseenheden in categorieën onderverdelen of classificeren. Wanneer we classificeren trachten we de elementen op basis van een bepaald kenmerk te sorteren en nemen we beslissingen over wie hetzelfde kenmerk heeft en wie hiervan verschilt. Het doel is ze zodanig in categorieën in te delen dat deze intern zoveel mogelijk homogeen zijn in vergelijking met de andere categorieën. We classificeren niet zomaar: de categorieën die we hanteren dienen om de sociale realiteit overzichtelijker te maken. Classificeren is fundamenteel in elke wetenschapstak. Alle andere niveaus van meting, hoe precies ook, bevatten inherent classificatie als minimaal onderdeel. Hierdoor kunnen we classificatie als het laagste meetniveau beschouwen. We geven arbitraire namen aan de categorieën als een soort van tekenset (tags, symbool) en maken geen veronderstellingen over de relatie tussen de categorieën. De volgorde waarin de categorieën worden opgesomd, impliceert in geen geval dat de ene categorie beter is dan de andere. Denk aan de categorie geslacht. Men is of man of vrouw, althans als we het biologische geslacht op basis van de meting als uitgangspunt nemen. Het lijkt eenvoudig, maar soms moet men toch twee keer nadenken. Neem het kenmerk nationaliteit: de meesten hebben één nationaliteit, maar sommigen hebben meerdere nationaliteiten. Hoe lossen we dit probleem op? We lossen het op door de definitie duidelijker af te bakenen. We zouden nationaliteit kunnen meten door te kijken naar de nationaliteit bij de geboorte, of door de nationaliteit van de vader of moeder te nemen. Je merkt hier dat er vervelende situaties ontstaan. Je moet een goede manier vinden

om iets te meten, anders is wat je meet compleet waardeloos. Er is bij nominale data geen sprake van een rangorde, ook al worden de categorieën voorgesteld door een getal. Denk bijvoorbeeld aan rugnummers van een elftal. De speler met rugnummer 14 is niet noodzakelijk beter dan de speler met rugnummer 7. Een classificatiesysteem dient aan twee belangrijke voorwaarden te voldoen: *exclusiviteit* en *exhaustiviteit*. Het principe van exclusiviteit houdt in dat de categorieën elkaar niet mogen overlappen. Een classificatiesysteem verliest zijn bruikbaarheid als de te classificeren data kunnen worden ondergebracht in meer dan één categorie. Dit betekent dat er slechts één criterium aan de basis van de opdeling mag liggen en dat dit criterium zo objectief mogelijk moet zijn. Het principe van exhaustiviteit impliceert dat het classificatiesysteem alle onderzoekseenheden moet kunnen classificeren, m.a.w.: elk geval moet in een categorie kunnen worden gebracht. Vandaar dat nominale classificatiesystemen vaak onderhevig zijn aan wijzigingen, doordat nieuwe categorieën ontstaan of worden bedacht. Zo zijn er bijvoorbeeld diverse vormen van criminaliteit die zijn ontstaan uit de opkomst van nieuwe technologieën (computercriminaliteit, gsm-fraude...) en die dus voordien gewoonweg niet bestonden. Of vormen van criminaliteit die voordien wel bestonden, maar die nu pas afzonderlijke aandacht krijgen en zodoende nu pas afzonderlijk worden benoemd (stalking, steaming, mobbing...). Door het feit dat de categorieën worden voorgesteld door arbitraire tekst of cijfers, kan men er, ook al worden de categorieën voorgesteld door cijfers, in geen geval echte statistische bewerkingen op uitvoeren. De functie van deze cijfers is immers enkel een label aan de categorie te geven. Bij de univariate statistiek houdt het vaak op met het beschrijven van de absolute en relatieve aantallen die zich in elke categorie bevinden.

Het **ordinale meetniveau** kenmerkt zich door het principe van de *ordenbaarheid* van de categorieën. Vaak is het mogelijk om categorieën te ordenen op basis van “de mate waarin ze een bepaald kenmerk” bevatten. Zo kan men mensen classificeren op basis van hun sociale status en grofweg drie categorieën onderscheiden: lage status, middelmatige status en hoge status. Iemand geklasseerd in de groep met een hoge status, geniet veel meer aanzien in de samenleving dan iemand met een lage status en kan op meer hulpbronnen beroep doen om problemen op te lossen dan iemand met een lage status. Bij een ordinale meting zijn we dus niet alleen in staat de mensen in categorieën in te delen (te classificeren) maar ook de categorieën onderling te *rangschikken of rangordenen*. Naast het aspect van onderscheidbaarheid, speelt m.a.w. nog een tweede aspect een rol: **ordening**. Hierdoor kunnen de categorieën op een **continuüm** worden geplaatst. Bij ordinale meetschalen zijn

zodoende steeds uitspraken mogelijk in de zin van ‘*meer of minder*’, ‘*groter of kleiner*’, ‘*beter of slechter*’. Om die reden spreekt men van ‘*geordend categoriserend meten*’. We kunnen alleen geen verschil maken in de sterkte van de relatie tussen de categorieën. We weten dus wel dat B groter is dan A maar kunnen niet aangeven hoeveel dit precies is. Het is dan ook zinloos om de tweede categorie van de eerste af te trekken of ze bij elkaar op te tellen. M.a.w., net als bij nominale variabelen kunnen we ook met ordinale variabelen geen wiskundige bewerkingen uitvoeren.

Variabelen van het nominale en ordinale meetniveau behoren beide tot de groep der **categorische variabelen**, ze zijn dus van het lagere meetniveau. Het is cruciaal deze terminologie te kennen, want deze komt heel vaak voor in de criminologie en in criminologisch onderzoek gebruikt men ze vaak. Deze variabelen kunnen **dichotoom** of **polytoom** zijn. Een *dichotomie* is een variabele die slechts twee waarden kan aannemen. Een *polytomie* is een variabele die meer dan twee categorieën kent. Een dichotome variabele kan slechts *twee mogelijke waarden* aannemen, bijvoorbeeld goed (1) of fout (0). De waarde 1 geeft aan dat de onderzoekseenheid het kenmerk heeft en de waarde 0 dat dit niet het geval is. Er bestaan variabelen die slechts twee waarden kunnen hebben, maar een variabele kan ook polytoom zijn, dit betekent dat de variabele *meerdere categorieën* heeft, bijvoorbeeld de *trichotomie*, de categorische variabele die drie categorieën aanneemt (bvb. 0= Geen slachtoffer geworden het afgelopen jaar, 1= 1 maal slachtoffer geworden het afgelopen jaar, 2= herhaald slachtoffer geworden het afgelopen jaar). Categorische variabelen kunnen worden gebruikt voor categorische data-analyse. Categorische data-analyse is het geheel van statistische analysetechnieken die beschikbaar zijn voor de statistische analyse van gegevens van het lage meetniveau.

### ***Interval meetniveau***

Als variabelen op **intervalniveau** gemeten zijn, is er niet alleen sprake van rangordening, maar hebben de *intervallen tussen de verschillende waarden* die een variabele aan kan nemen ook een **exakte betekenis**. Een gekend voorbeeld is temperatuur. Het verschil tussen 5 en 10 °C is even groot als het verschil tussen 10 en 15 °C (5 °C). Er is sprake van een *vaste meeteenheid* waarbij de waarden voor de graden betekenis toekennen aan de afstanden tussen de graden. Wat je echter niet kunt zeggen is dat 20 °C twee keer zo warm is als 10 °C. Dit komt door het ***ontbreken van een natuurlijk (of absoluut) nulpunt***. Het nulpunt bij graden Celsius is namelijk arbitrair. Er zijn meer manieren om temperatuur te meten, zoals door

middel van graden Fahrenheit. Bij meting in graden Fahrenheit is er een ander nulpunt en de intervallen tussen de graden zijn anders dan bij graden Celsius. Wanneer het in Amerika tien graden warmer wordt, is die temperatuur in de regel gemeten in Fahrenheit. Wanneer in België de temperatuur met tien graden stijgt, is dit niet dezelfde warmtestijging, omdat wij hier in graden Celsius rekenen.

Variabelen als inkomensklassen en leeftijdsgroepen kun je ook op intervalniveau meten als je ervoor zorgt dat de *afstanden tussen de waarden altijd even groot zijn*. Stel, je kiest de waarden voor de variabele leeftijdsgroepen als volgt:

Klasse 1: 21 - 25 jaar;

Klasse 2: 26 - 30 jaar;

Klasse 3: 31 - 35 jaar;

Klasse 4: 36 - 40 jaar;

...

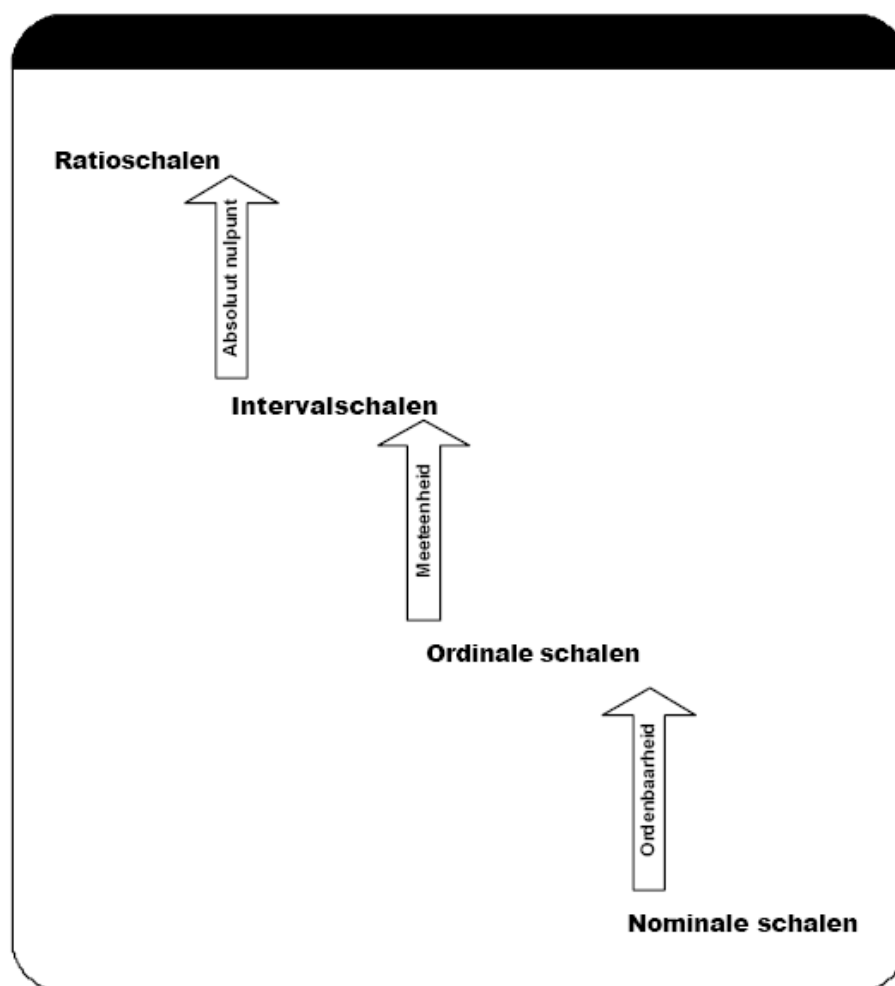
De afstanden tussen de waarden zijn in dit voorbeeld steeds even groot. De intervallen hebben daarmee een betekenis. Je kunt zeggen dat het verschil tussen klasse 3 en klasse 4 net zo groot is als het verschil tussen klasse 2 en 3. En er is geen absoluut of natuurlijk nulpunt. Je kunt niet zeggen dat iemand uit groep 4 vier keer zo oud is als iemand uit groep 1. Met deze indeling van de leeftijdsgroepen hebben we dus een variabele die op intervalniveau gemeten is. Dit in tegenstelling tot bijvoorbeeld de waarderingsschaal ‘zeer slecht’, ‘slecht’, ‘middelmatic’, ‘goed’ en ‘zeer goed’. We kunnen hier niet zeggen dat het verschil tussen zeer slecht en slecht even groot is als het verschil tussen goed en zeer goed, noch kunnen we dit verschil uitdrukken in aantal meeteenheden. Slechts weinig begrippen in de sociale wetenschappen zijn metrisch van aard. Voor de meting van intelligentie, sociale status, onveiligheidsgevoelens, vooroordelen,... bestaan er geen vaste eenheden waarover alle sociale wetenschappers het eens zijn en die, wanneer ze herhaaldelijk weer worden gemeten, steeds dezelfde waarden opleveren.

### ***Ratio meetniveau***

Het niveau waarbij sprake is van *rangordening* en waarbij de *intervallen betekenis* hebben en er een *natuurlijk nulpunt* aanwezig is, is **het rationiveau**. Op dit niveau is nul ook werkelijk een absoluut, niet-arbitrair nulpunt. Te denken valt aan lengte, gemeten in aantal centimeter. Er is dan een absoluut nulpunt: iets met een lengte van 0,000 centimeter heeft geen lengte, gemeten in centimeter. Misschien vind je dit verwarrend, want je kan denken: lengte kan toch

ook in millimeter worden uitgedrukt? En in de kwantummechanica bestuderen we de micro-wereld die helemaal niet zichtbaar is met het blote oog. Dat is allemaal juist, maar je mag je niet laten afleiden door zulke veronderstellingen. Het gaat om het principe dat een kenmerk, gemeten op een bepaalde manier geen negatieve waarden kan aannemen. Nu hebben niet alleen de verschillen tussen de afzonderlijke waarden betekenis, maar ook het quotiënt (het resultaat van een deling). Een krantenartikel met een kolomlengte van 15 centimeter is drie keer langer dan een krantenartikel met een kolomlengte van 5 centimeter. En het verschil in kolomlengte tussen deze twee krantenartikelen is 10 centimeter. Andere voorbeelden van een ratio meetniveau zijn leeftijd in jaren, tijd in minuten, prijs, het exacte inkomen in euro's per maand.

**Schematisch kunnen we de hiërarchie in de verschillende meetniveaus als volgt voorstellen:**



Deze hiërarchie kan ook in een tabel worden samengevat:

Meetniveau	Classificatie	Totale ordening	meeteenheid	Absoluut nulpunt
Nominaal	Ja	<i>Neen</i>	<i>Neen</i>	<i>Neen</i>
Ordinaal	Ja	Ja	<i>Neen</i>	<i>Neen</i>
Interval	Ja	Ja	Ja	<i>Neen</i>
Ratio	Ja	Ja	Ja	Ja

We willen je hier een belangrijke tip meegeven. **De vier onderscheiden meetniveaus zijn cruciaal in de statistiek.** Ze bepalen immers de wijze waarop we de gegevens grafisch kunnen voorstellen, de parameters die mogen gehanteerd worden en de analysetechnieken die we mogen gebruiken. Eén van de meest voorkomende fouten in de statistiek heeft te maken met het zondigen tegen deze meetniveaus. In principe trachten we bij metingen een zo hoog mogelijk meetniveau te verkrijgen op voorwaarde dat dit inhoudelijk zinvol en mogelijk is. Het is immers voor heel wat kenmerken onmogelijk om het meetniveau te verhogen. Nationaliteit en geslacht bijvoorbeeld, zijn altijd nominaal. Soms is het onmogelijk om een meting op een hoger meetniveau te tillen door weerstand van de respondent tegen een té precies antwoord (bv. inkomen: men geeft het liever in een interval dan het juiste bedrag te noemen; seksueel gedrag), soms is de herinnering onprecies en dus zinloos zeer precies te bevragen (hoe vaak heeft u afgelopen jaar persoonlijk een politieagent gesproken?).

## 7. Discrete en continue variabelen

Bij **continue meetschalen** kunnen *alle mogelijke waarden* de uitkomst zijn van de meetprocedure: niet alleen bijvoorbeeld de waarden 1 en 2, maar ook 1,2 en 1,75. Dit in tegenstelling tot een **discrete meetschaal**, die *beperkt is tot een telbaar aantal waarden*, waarvoor je in de regel gehele getallen gebruikt. De tussenliggende waarden hebben bij discrete meetschalen geen betekenis (mensen hebben 1, 2 of 3 televisies in huis, maar geen 1,5 televisie). Een continu verschijnsel wordt vaak toch met een discrete meetschaal gemeten en dan hebben tussenliggende waarden wel een betekenis. De onderzoeker heeft dan de keuze gemaakt een kenmerk met een beperkte precisie te meten, bijvoorbeeld het meten van leeftijd in jaren en niet in jaren plus aantal maanden, of nog preciezer in aantal dagen. Voorbeelden van continue verschijnselen zijn (leef)tijd en afstanden. Bijvoorbeeld, bij tijd kun je naast een onderscheid in uren een onderscheid maken in minuten, seconden, of zelfs nanoseconden.

Ook de tussenliggende waarden hebben dan een betekenis. Voorbeelden van discrete verschijnselen zijn het aantal kinderen in een gezin of het aantal televisietoestellen in een huis. Je hebt niet 1,5 televisietoestel of een half kind.

## 8. De datamatrix als input voor statistische analyses

Om statistische analyses te kunnen uitvoeren, dienen we eerst gegevens te verzamelen. Deze gegevens worden in een *datamatrix* (*gegevensmatrix*) geplaatst, zodat ze verder geanalyseerd kunnen worden. Een gegevensmatrix bevat de informatie van elke statistische eenheid waarover men informatie heeft verzameld. Die informatie heeft betrekking op kenmerken die variëren (variabelen) en de statistische eenheden nemen dus verschillende waarden aan op deze kenmerken. Een gegevensmatrix krijgt de vorm van een  $R \times K$  tabel, i.e. bevat informatie in *rijen* ( $R$ ) en *kolommen* ( $K$ ). *De conventie wil dat we in de rijen de statistische eenheden plaatsen en in de kolommen de kenmerken van deze personen.* In het voorbeeld hieronder tonen we enkele respondenten met hun kenmerken uit een bevraging. De rijen worden gevormd door de respondenten met elk hun respondentnummer. De kolommen worden gevormd door kenmerken van deze respondenten of de variabelen. Een datamatrix wordt vaak opgesteld in een statistisch verwerkingspakket zoals SPSS, maar kan ook in een werkblad zoals Excel worden opgemaakt.

**Een ruwe datamatrix ziet er als volgt uit:**

Resnr	Naam	Geslacht	opleiding	diploma vader	#bedreigd	leeftijd
1	Jan	man	menswet.	LO	12	18
2	An	vrouw	lat.wisk	HO	5	17
3	Piet	man	mod.talen	LO	35	19
4	Els	vrouw	economie	HO	0	18

Vaak worden numerieke codes gebruikt om kenmerken van respondenten voor te stellen. Dit wordt gedaan in wat we noemen een “**codeboek**”, of een lijst waarin uitgelegd staat welke inhoudelijke betekenis wordt gegeven aan de numerieke code. Deze codes moeten achteraf in statistische verwerkingspakketten worden ingegeven zodat we niet de code maar het bijbehorende label in de grafieken en tabellen die we maken, te zien krijgen. In het voorbeeld wordt de numerieke code “0” gegeven aan vrouwelijke respondenten en de numerieke code



“1” aan mannen. Op een analoge manier werden hier codes gegeven aan de variabelen opleiding en diploma vader. We geven die codes omdat het makkelijker is om codes in te voeren dan telkens de naam van een categorie voluit te schrijven.

**Een datamatrix gebaseerd op een codeboek, ziet er als volgt uit:**

Resnr	geslacht	opleiding	diploma vader	#bedreigd	leeftijd
1	0	0	0	12	18
2	1	1	1	5	17
3	0	2	0	35	19
4	1	3	1	0	18

## **9. Een handige afrondingsregel voor statistische gegevens**

Meten wordt helaas te vaak voorgesteld als een zeer precies proces dat dé juiste waarde oplevert. In de exacte wetenschappen, waar het meten aan de hand van duidelijke criteria zoals minuten, lengte, afstanden, snelheden,... de normaalste gang van zaken is, is men doordrongen van het feit dat er steeds meetfouten ontstaan. In de criminologie, waar het meten veel minder aan de hand van duidelijke en onbetwistbare maatstaven gebeurt, schijnt men het probleem van meetfouten nauwelijks in te zien. Het argument is vaak dat er in de sociale wetenschappen al zoveel andere problemen zijn om bepaalde zaken meetbaar te maken dat de statistische meting zelf niet zo problematisch is: het maakt niet uit hoe fijn we trachten te meten. Dat is althans de idee die sommige criminologen hebben en dit idee is fout en getuigt van weinig inzicht in de statistiek en de effecten ervan op resultaten. Het zou ons te ver brengen om grondig op dit probleem in te gaan. We werpen het hier enkel op om het probleem van afronden te behandelen. Aangezien we steeds meetfouten maken, is het weinig zinvol om de metingen in meer dan twee cijfers na de komma uit te drukken. Het wekt de indruk van een (te) grote precisie bij het meten, maar weerspiegelt in geen geval de bestaande meetpraktijken. We gebruiken dus in dit handboek twee cijfers na de komma. We hadden kunnen opteren voor drie, maar het principe blijft hetzelfde.

Een regel die bij het afronden vaak gehanteerd wordt is:

**Men rondt naar beneden af van 1 tot 4**

**Men rondt naar boven af van 5 tot 9**

**We kijken steeds enkel naar het eerstvolgende decimaal om af te ronden.**

Een voorbeeld van afronden:

- 5,441 – 5,442 – 5,443 – 5,444 wordt 5,44
- 5,445 wordt 5,45
- 5,446 – 5,447 – 5,448 – 5,449 wordt 5,45
- 5,431 – 5,432 – 5,433 – 5,434 wordt 5,43
- 5,4356 wordt 5,44
- 5,436 – 5,437 – 5,438 – 5,439 wordt 5,44

## 10. Het sommatieteken

Iedereen weet wel wat een som is, met name het optellen van cijfers. Denk maar aan de eenvoudige som  $2+2$ . Echter, in de geavanceerde wiskunde kunnen zulke sommen uit ettelijke elementen bestaan. Daarom bestaat er een compacte en effectieve manier om een sommatie aan te geven. Dit wordt in de wiskunde gedaan met  $\Sigma$ , de hoofdletter sigma uit het Griekse alfabet.

Praktisch toegepast:

$$\sum_{i=m}^n xi = x_m + x_{m+1} + x_{m+2} + x_{m+3} + \dots + x_{n-2} + x_{n-1} + x_n$$

De letter  $i$  duidt de *index* aan, maar zou evengoed  $k$  of een andere letter kunnen worden genoemd. Het getal  $m$  is de *ondergrens van de sommatie*, en het getal  $n$  de *bovengrens van de sommatie*. De vermelding van  $i=m$  geeft aan dat de sommatie begint met de term met als index de waarde  $m$ . Elke volgende term heeft een index 1 hoger dan de vorige, en de sommatie eindigt met de term met index  $n$ .

Een uitgewerkt voorbeeld vinden we hieronder:

$$\sum_{i=1}^5 2^i = 2^1 + 2^2 + 2^3 + 2^4 + 2^5 = 2 + 4 + 8 + 16 + 32 = 62$$

Er staat neem de som van “twee tot de macht  $i$ , waarbij  $i$  van één tot vijf varieert”.

### 11. Afspraken bij het presenteren van tabellen

In de loop van deze cursus komen we verschillende wijzen tegen om een datamatrix samen te vatten in tabellen. We onderscheiden:

- a. **Frequentietabellen:** tellingen van hoe vaak iedere waarde van een variabele voorkomt.
- b. **Kruistabellen:** tellingen van hoe vaak waarden van twee variabelen in combinatie met elkaar voorkomen.

Het doel hiervan is om de gegevens overzichtelijk te presenteren. Op die manier kunnen we algemene uitspraken doen over criminologisch relevante onderwerpen. Als we ruwe gegevens samenvoegen, gaat een deel ervan verloren, maar dat is niet zo erg: het grote voordeel is dat dit ten goede komt aan de overzichtelijkheid. Laten we daarom bij aanvang van deze cursus een aantal afspraken maken omtrent het presenteren van tabellen:

Een tabel heeft **altijd**:

- a. Een titel
- b. Een bronvermelding onderaan de gegevens
- c. Een vermelding van de waarnemingseenheid
- d. Een vermelding van de variabelen
- e. Een vermelding van de meeteenheid
- f. Een logische en overzichtelijke indeling

Hiertegen wordt soms gezondigd wanneer studenten een paper schrijven. Correct is het echter niet als je deze regels negeert. We wijzen er op dat het belangrijk is de regels te volgen en een tabel steeds consequent te presenteren. Beschouw deze regels als de “grammatica” van de kwantitatieve criminoloog. Net zoals de gewone talen, hebben formele talen zoals de statistiek, hun regels. Als we deze volgen, kunnen we elkaar begrijpen. Als we dat niet doen, creëren we chaos.

## **12. Leerdoelen**

Op het einde van dit hoofdstuk dienen volgende zaken gekend te zijn:

Je kent alle begrippen die we in dit hoofdstuk hebben geïntroduceerd. Je weet het onderscheid tussen univariate, bivariate en multivariate beschrijvende statistiek en je weet wat inferentiële statistiek voor ogen houdt. Je kent het onderscheid tussen onderzoekseenheden en variabelen. Je hebt inzicht in de verschillende meetniveaus die we in de statistiek hanteren en weet ook dat die meetniveaus een hiërarchische structuur in zich hebben. Je weet wat een gegevensbestand of datamatrix is en kan de scores van respondenten (de onderzoekseenheden) aflezen uit zo een tabel. Je weet hoe een tabel er dient uit te zien.