

**The Gluecks' delinquent sample: actual mean number of offences for total crime (total events = 9,548): ages 7 to 70**

Source: Laub and Sampson (2003), *Shared Beginnings, Divergent Lives*, Fig 5.21, p. 86.

# BASISCURSUS STATISTIEK IN DE CRIMINOLOGIE

## Deel I

## THEORIEBOEK

Prof. dr. Lieven Pauwels



Copyright © 2025: Lieven Pauwels

Auteur: Lieven Pauwels

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of enige andere manier, zonder voorafgaande schriftelijke toestemming van de rechthebbende.

Foto cover: Lambert Adolphe Jacques Quetelet (Gent, 22 februari 1796 – Brussel, 17 februari 1874)

## **Voorwoord**

In criminologisch onderzoek nemen de opzet van een bevolkingsbevraging of een experiment en het uitwerken van de onderzoeksgegevens een belangrijke plaats in. Welke rol speelt de statistiek hierin? Met behulp van statistische methoden kunnen de verkregen criminologisch relevante gegevens overzichtelijk gemaakt worden door deze te beschrijven (beschrijvende statistiek) en kunnen conclusies getrokken worden over het veel grotere geheel (we spreken van bevolking of populatie) waaruit de waarnemingen (we spreken van de steekproef) afkomstig zijn, dit door middel van statistisch toetsen (testen van veronderstellingen) en schatten van onzekere uitkomsten in de bevolking op basis van informatie uit een steekproef (inferentiële statistiek). Voor het toetsen en schatten is enige achtergrondkennis van de belangrijkste kansverdelingen noodzakelijk. Met deze cursus leren we jullie om te herkennen welke statistische methoden bij de beantwoording van criminologisch relevante probleemstellingen gebruikt kunnen worden. De rode draad doorheen deze cursus statistiek is de volgende: je verliest je onderzoeksraag nooit uit het oog (wat wil je precies weten) en je kiest steeds een verantwoorde statistische methode die past bij de vraag en bij de aard van het kenmerk of de kenmerken die centraal staan in je onderzoeksraag.

De cursus bestaat uit hoorcolleges en werkcolleges. Tijdens de hoorcolleges wordt de rode draad van de statistische theorie uiteengezet. Daarna verwerk je in een werkcollege de theorie. De opgaven hebben vooral het analyseren en kritisch interpreteren van criminologische onderzoeksresultaten tot doel. Het rekenwerk wordt tot een minimum herleid en is enkel gericht op het begrijpen van wat er gebeurt achter de schermen van een softwarepakket dat complexe analyses op grote gegevensbestanden uitvoert in een fractie van een seconde. We stellen gegevens uit reëel gevoerd criminologisch onderzoek ter beschikking van de studenten om de link met de statistische praktijk zo duidelijk mogelijk te maken. De statistiek is wijder verspreid in het criminologisch veld dan je op het eerste zicht zou denken. De hoorcolleges bevatten niet enkel contacturen, maar bevatten flink wat begeleide zelfstudie ter voorbereiding van de werkcolleges. Met begeleide zelfstudie bedoelen we dat je zelf een aantal vaardigheden thuis dient te verwerven en dat we deze

daarna via de hoorcolleges en werkcolleges doornemen en bespreken. Er wordt dus van je verwacht dat je thuis (tijdens de zelfstudie) de theorie en de opgaven voorbereidt en (nog eens) bestudeert. Voor de zelfstudie heb je ook de beschikking over het statistisch verwerkingspakket SPSS, dat je via Athena kan hanteren.

Wil je slagen voor dit opleidingsonderdeel dan kunnen we enkel aanbevelen dat je actief meewerkt in de werkcolleges. De klemtoon in de opleiding criminologie komt te liggen op de interpretatie, maar in een eerste jaar vinden we het nog steeds belangrijk dat je een minimale bagage hebt van de achtergrond. Tegenwoordig wordt de meest complexe analyse via verwerkingspakketten op een seconde uitgerekend. De criminoloog is echter niet zomaar iemand die op een knopje drukt, maar weet waarom via een druk op de knop een analyse wordt uitgevoerd. In latere cursussen wordt voortgebouwd op de kennis uit dit handboek en zullen geavanceerde technieken exclusief aan de hand van SPSS worden aangeleerd.

**Tot slot nog een laatste opmerking over het instuderen van statistiek. Dit is de grootste valkuil waar de student in trapt. Statistiek is beslist niet iets wat je het laatste weekend voor het examen nog wel even leert. Het is een kwestie van regelmatig met het opleidingsonderdeel bezig te zijn, thuis de theorie voorbereiden en veel opgaven maken om zodoende de theorie te verwerken en je de statistische manier van denken eigen te maken!**

Wij wensen je daarbij heel veel succes toe!

Prof. dr. Lieven Pauwels

**Verantwoordelijk docent**

**Prof. dr. Lieven J.R. Pauwels**

**Vakgroep Criminologie, Strafrecht en Sociaal Recht**

**Directeur Institute for International Research and Criminal Policy (IRCP)**

**Universiteitstraat 4-6**

**9000 Gent**

**Tel: ++32 (0)9 264 68 37**

**[Lieven.Pauwels@ugent.be](mailto:Lieven.Pauwels@ugent.be)**



[www.fokksuk.nl](http://www.fokksuk.nl)



## Inhoudstafel

Voorwoord .....	i
-----------------	---

### **Hoofdstuk 1: De logica van statistische vergelijkingen en analyses .....1**

1. Inleiding: waarom data analyseren? .....	1
---	---

2. Geschiedenis van de statistiek in een notendop .....	2
---	---

3. Het gebruik van statistiek.....	3
------------------------------------	---

4. Theorieconstructie in een oogopslag .....	5
--	---

<i>Wat is theorie?</i> .....	5
------------------------------	---

<i>Theorie en onderzoek</i> .....	6
-----------------------------------	---

5. Het proces van wetenschappelijk onderzoek .....	7
--	---

<i>Observatie en nieuwsgierigheid</i> .....	9
---	---

<i>Centrale onderzoeks vragen</i> .....	10
---	----

<i>Onderzoeksdeelvragen</i> .....	10
-----------------------------------	----

6. Onderzoek: bewegen van theorie naar data en terug .....	11
--	----

<i>Hypothesen formuleren</i> .....	12
------------------------------------	----

<i>Constructie van het onderzoeksdesign</i> .....	13
---	----

<i>Conceptualisering</i> .....	13
--------------------------------	----

<i>Operationalisering</i> .....	13
---------------------------------	----

<i>Data verzamelen</i> .....	14
------------------------------	----

<i>Conclusies trekken</i> .....	15
---------------------------------	----

<i>Communiceren van resultaten</i> .....	16
--	----

### **Hoofdstuk 2: Inleidende begrippen.....18**

1. Inleiding .....	18
--------------------	----

2. Beschrijven, schatten en veralgemenen als statistische bedrijvigheid .....	20
---	----

3. Statistiek en de beantwoording van beschrijvende en verklarende onderzoeks vragen .....	22
--	----

4. Statistische eenheden .....	24
5. Univariate, bivariate en multivariate beschrijvende analyse .....	26
6. Meetniveaus van variabelen .....	27
<i>Het nominale meetniveau en het ordinale meetniveau .....</i>	27
<i>Interval meetniveau.....</i>	29
<i>Ratio meetniveau.....</i>	30
7. Discrete en continue variabelen .....	32
8. De datamatrix als input voor statistische analyses.....	33
9. Een handige afrondingsregel voor statistische gegevens.....	34
10. Het sommatieteken.....	35
11. Afspraken bij het presenteren van tabellen.....	36
12. Leerdoelen.....	37

<b>Hoofdstuk 3: De univariate beschrijvende statistiek.....</b>	<b>38</b>
1. Inleiding .....	38
2. Over absolute en relatieve frequenties en hun grafische voorstelling .....	38
<i>Grafische voorstellingen.....</i>	43
<i>Taartdiagram of cirkelgrafiek (pie chart).....</i>	44
<i>Staafdiagram (bar chart).....</i>	45
<i>Cumulatief frequentiediagram.....</i>	46
<i>Histogram.....</i>	47
<i>Lijndiagram.....</i>	49
<i>Frequentiepolygoon.....</i>	49
<i>Opgelet met grafische voorstellingen.....</i>	50
3. Parameters van centraliteit.....	51
<i>De modus .....</i>	52
<i>De mediaan .....</i>	52

<i>De kwantielen</i> .....	53
<i>Het rekenkundig gemiddelde</i> .....	54
<i>Verantwoord kiezen tussen centrummaten</i> .....	56
4. Parameters van spreiding: vive la différence!.....	57
<i>De variatieratio (VR)</i> .....	58
<i>De index van diversiteit (ID)</i> .....	58
<i>De variatiebreedte</i> .....	59
<i>De interkwartielafstand (K3-K1)</i> .....	60
<i>Spreidingsmaten op metrisch niveau</i> .....	60
<i>De gemiddelde absolute afwijking</i> .....	62
<i>De variatie</i> .....	62
<i>De (steekproef)variantie</i> .....	62
<i>De (steekproef)standaardafwijking</i> .....	63
5. Zelf uitrekenen van gemiddelde, variantie en standaardafwijking .....	63
<i>De variatiecoëfficiënt</i> .....	65
6. Parameters van vorm.....	66
7. De Box-plot.....	69
8. Testvragen.....	78
8. Leerdoelen.....	83
<b>Hoofdstuk 4: Een inleiding in kansrekenen.....</b>	<b>87</b>
1. Waarom kansrekenen voor criminologen? .....	87
2. Kansdefinities .....	88
3. Kansregels .....	90
4. Permutaties en combinaties.....	93
5. Kansvariabelen en de binomiale verdeling .....	94
6. De binomiale verdeling.....	96

7. De binomiale verdeling gaat over in een normale verdeling .....	99
8. Waarom is de binomiale verdeling belangrijk in kwantitatief criminologisch onderzoek?.....	102
9. Leerdoelen.....	103

## **Hoofdstuk 5: De standaardnormale verdeling en diens eigenschappen .....105**

1. Inleiding .....	105
2. De normale en standaardnormale verdeling .....	106
3. Van normale verdeling naar standaardnormale verdeling .....	109
4. Z-scores en het gebruik van de tabel van de standaardnormale verdeling .....	109
5. Leerdoelen.....	113

## **Hoofdstuk 6: Inleiding tot de bivariate beschrijvende statistiek.....115**

1. Inleiding: causale relaties versus statistische relaties .....	115
2. Causaliteit op een bierviltje .....	117
3. Symmetrische en asymmetrische relaties tussen variabelen.....	119
4. Doelstelling van de bivariate beschrijvende statistiek .....	121
5. Bivariate frequentieverdelingen voor lage en hoge meetniveaus .....	123
6. Verantwoord kiezen tussen een reeks van associatiematen .....	127
7. Leerdoelen.....	127

## **Hoofdstuk 7: Bivariate associatiematen voor nominale en ordinale variabelen.....129**

1. Inleiding .....	129
2. Het percentageverschil als associatiemaat op nominaal niveau .....	129
3. De odds ratio als associatiemaat op nominaal niveau.....	133
4. Chi-kwadraat ( $\chi^2$ ) als associatiemaat op nominaal niveau.....	136
5. Phi .....	141
6. Cramer's V .....	141

7. Gamma als associatiemaat op ordinaal niveau .....	142
8. De rangcorrelatiecoëfficiënt van Spearman en Kendall's Tau-b.....	144
9. Leerdoelen.....	146

## **Hoofdstuk 8: Correlatie- en regressieanalyse.....147**

1. Symmetrische associatiematen voor kenmerken op metrisch niveau.....	147
<i>De covariatie.....</i>	153
<i>De covariantie.....</i>	154
<i>De product-moment correlatiecoëfficiënt van Pearson.....</i>	155
2. Covariatie, covariantie en correlatie: een uitgewerkt rekenvoorbeeld.....	156
<i>Stappen te volgen in het uitrekenen van een correlatie .....</i>	157
3. De bivariate lineaire regressieanalyse als asymmetrische analysetechniek.....	158
4. Zelf uitrekenen van de parameters van de regressierechte .....	169
<i>Stappen te volgen in het uitrekenen van een bivariate regressie.....</i>	170
5. De rapportage van de belangrijkste parameters van de regressierechte in een rapport .....	173
6. En wat als de meetniveaus van twee variabelen verschillend zijn?.....	174
7. Leerdoelen.....	175

## **Hoofdstuk 9: Inferentiële statistiek en variantieanalyse .....177**

1. Waarom gebruiken we inferentiële statistiek? .....	177
2. De representativiteit van steekproeven .....	178
3. Steekproeven en populatie .....	180
4. Steekproeven en het principe van toeval .....	181
5. De theorie van toevalssteekproeven.....	182
6. Kenmerken van steekproevenverdelingen .....	185
7. Het gebruik van de normale verdeling in de inferentiële statistiek .....	187
8. De centrale limietstelling .....	187

9. Puntschatting en intervalschatting .....	189
10. Het berekenen van een betrouwbaarheidsinterval rond een parameter .....	193
11. Statistische hypothesetoetsing .....	195
12. Eenzijdig of tweezijdig toetsen van een nulhypothese? .....	200
13. Andere belangrijke verdelingen.....	201
14. De variantieanalyse als toets voor verschillen tussen groepen inzake metrische kenmerken.....	203
15. Zelf uitrekenen van een variantieanalyse.....	205
16. Voorbeelden van statistische inferentie in andere analysetechnieken .....	210
17. Testvragen.....	213
18. Leerdoelen.....	222

**Hoofdstuk 10: De partiële correlatie als introductie tot de multivariate statistiek .....225**

1. Inleiding .....	225
2. De partiële correlatiecoëfficiënt .....	227
3. De berekening van de partiële correlatiecoëfficiënt a.h.v. regressievergelijkingen .....	233
4. Berekening van de partiële correlatiecoëfficiënt a.h.v. rekenkundige formule .....	241
5. Suppressie-effect.....	243
6. Leerdoelen.....	244

**Hoofdstuk 11: Regressieanalyse met twee onafhankelijke variabelen.....245**

1. Inleiding .....	245
2. De noodzaak voor het meten van controlevariabelen .....	246
3. De vergelijking tussen twee bivariate versus één meervoudige regressie .....	248
4. De uitbreiding naar een meervoudige regressieanalyse.....	250
5. Het relatieve belang van elke onafhankelijke variabele .....	251
6. De berekening van de gestandaardiseerde gewichten ( $\beta_1$ en $\beta_2$ ) .....	253

7. Veronderstellingen bij het uitvoeren van een lineaire regressie analyse.....	255
8. Controle op de regressievoorwaarden.....	259
<i>Normaliteit</i> .....	259
<i>Heteroscedasticiteit</i> .....	259
<i>Additiviteit</i> .....	260
<i>Lineariteit</i> .....	261
<i>Uitbijters of outliers</i> .....	261
9. De limieten van meervoudige regressie .....	262
10. Testvragen .....	263
11. Leerdoelen.....	265
<b>Hoofdstuk 12: Complexere relaties tussen variabelen .....</b>	<b>267</b>
1. Inleidende begrippen.....	267
2. Mediatorvariabele .....	267
3. Moderatorvariabele of het interactie-effect.....	268
4. De pad-analyse.....	282
<i>Directe en indirecte effecten</i> .....	283
5. De berekening van de totale en indirecte effecten in de pad-analyse .....	284
6. Nog een voorbeeld van een pad-model.....	285
7. Een rekenvoorbeeld op basis van de gestandaardiseerde padcoëfficiënten.....	290
8. Leerdoelen.....	292
<b>Slotbeschouwingen.....</b>	<b>295</b>
<b>Oplossingen testvragen.....</b>	<b>297</b>
<b>Synthese-oefening.....</b>	<b>313</b>
<b>Referenties .....</b>	<b>317</b>
<b>Bijlage1: Tabellen van statistische verdelingen</b>	



## **Hoofdstuk 1**

### **De logica van statistische vergelijkingen en analyses**

*'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.'*

**-H. G. Wells**

#### **1. Inleiding: waarom data analyseren?**

Ontdekking en vernieuwing zijn allicht de grote verschillen tussen de moderne bedrijvigheid en die van onze voorvaderen. Sinds de Renaissance is er een nadruk komen te liggen op het leren en verbeteren van de wijze waarop we dingen doen. Wetenschappers, uitvinders en anderen die betrokken zijn in het proces van wetenschappelijk onderzoek, hebben in het verleden vaak eerbied gekregen voor hun werk. Zo worden in talloze schoolboeken onder meer Galileo, Newton, Einstein en Madam Curie nog steeds geprezen voor hun werk en uitvindingen. Zoals jullie zullen leren in deze cursus, zou dit ook zo moeten zijn voor mensen zoals Pearson en Kendall. En voor onze eigen Adolphe Quetelet, die de normaalverdeling ontdekte voor menselijke eigenschappen, zich bezig hield met statistiek als hulpmiddel om de moraliteit, de zeden en gewoonten van gebieden te bestuderen en op basis daarvan beleidsbeslissingen te nemen. Quetelet wist al heel vroeg dat mensen verschilden in de mate waarin zij betrokken zijn bij regelovertredend gedrag en dat gebieden verschilden in de mate waarin zij concentraties kenden van criminaliteit.

Wetenschap gaat onder meer over het **ontdekken van patronen en processen**. Veel patronen komen niet zomaar uit de lucht gevallen maar zijn het gevolg van het feit dat mensen nu eenmaal gewoontedieren zijn en een menselijke natuur hebben, die in samenspraak met culturele omstandigheden, attitudes, denkprocessen, morele emoties, overtuigingen over wat goed en verkeerd is, ... beïnvloeden. Het proces van **wetenschappelijk onderzoek** voorziet in een methode om zaken die ons op een systematische manier interesseren, te bestuderen. In het algemeen vereist dit proces bewijs om een argument te ondersteunen. Eén van de duidelijkste methoden om bewijs

te verkrijgen is door kwantitatieve data te bestuderen die geassocieerd zijn met objecten die bestudeerd worden. Dit gebeurt door statistische analyse.

## **2. Geschiedenis van de statistiek in een notendop**

De vroegste vorm van wat nu gezien wordt als **statistische analyse** werd ontwikkeld door Pythagoras in de 6<sup>de</sup> eeuw voor Christus (het gemiddelde). Dit was de voorloper van de **beschrijvende statistieken**. Het andere type van statistische analyse (**inferentiële statistiek**) werd -waarschijnlijk- voor het eerst ontwikkeld in het Oosten rond 200 voor Christus (Dudycha & Dudycha, 1972). Dit was in de vorm van een waarschijnlijkheidsanalyse (*probability analysis*) om na te gaan of een verwachte baby eerder een jongen of meisje zou zijn. Waarschijnlijkheidstheorie/Kansberekening (*Probability theory*), zoals het later zou noemen, evolueerde verder in de vorm van gokberekeningen (*gambling mathematics*) in het werk van Blaise Pascal (1623-1662) en Christiaan Huygens (1629-1695) (David, 1962). In de late jaren 1800 en vroege jaren 1900 werden nog vele andere beschrijvende statistieken ontwikkeld door mathematici en wetenschappers zoals Sir Francis Galton (1883), een neef van Charles Darwin, en Karl Pearson (1895).

Statistiek ging verder dan gokken en pure mathematische concepten en werd later '*political arithmetics*' genoemd. Een term dat te danken is aan de dichte associatie met diegenen die politieke onderwerpen bestudeerden, inclusief economische. Men probeerde immers op basis van gegevens met behulp van '*political arithmetics*' antwoorden op problemen van de staat te krijgen. Hier ontstond waarschijnlijk de associatie tussen leugentjes bedenken om er politiek voordeel uit te halen en liegen door statistieken te manipuleren. Inderdaad, statistieken worden wel eens gemanipuleerd door politici, maar dat is geen reden om de statistiek af te schrijven. Dat zou veeleer een reden moeten zijn om de leugenachtige politicus af te schrijven. Statistiek is vaak de pianist waarop geschoten wordt. Als we de overleveringen mogen geloven dan was de eerste politicus die van statistiek gebruik maakte een zekere John Graunt (1662). Hij gebruikte beschrijvende statistieken, zoals we ze vandaag noemen, om het sterftecijfer in Londen te bestuderen. Hoewel er een hevig debat gaande is aangaande het originele gebruik van de term '**statistiek**' of '**statistics**' (Yule, 1905), is het merendeel van de statistici het erover eens dat de term statistiek uitgevonden

is door **Eberhard August Wilhelm von Zimmerman** in de inleiding van ‘*A Political Survey of the Present State of Europe*’ (1787).

Het moderne gebruik van de term *statistiek* (ten opzichte van *wiskunde*) wordt vaak toegeschreven aan R. A. Fisher en zijn werk ‘*Statistical Methods for Research Workers*’ (1925), waarin hij stelt dat “*a statistic is a value calculated from an observed sample with a view to characterizing the population from which it is drawn.*” Sinds die tijd hebben vele statistici bijgedragen tot de technieken die mogelijk zijn om data te analyseren. Veel procedures hebben ook de naam gekregen van deze statistici. De bijdrage aan statistische technieken gebeurt vandaag de dag nog steeds. Analysetechnieken zijn de laatste jaren enorm toegenomen en het heeft onderzoekers in de sociologie, criminologie en andere wetenschappen mogelijk gemaakt om de relatie tussen variabelen accurater te bestuderen.

### **3. Het gebruik van statistiek**

De term ‘statistiek’ wordt vaak verkeerd begrepen omdat deze term eigenlijk twee praktische toepassingen kent. De eerste term, die de historiek van de term reflecteert, is de verzameling en bewaring van data, vaak uitgedrukt in een samenvattende vorm. Goede voorbeelden hiervan zijn de volkstelling- of sterftestatistieken. Volkstellingstatistieken geven een voorstelling van de karakteristieken of kenmerken van de personen die in een land wonen op een bepaald moment in de tijd en sterftestatistieken geven inzicht in het aantal personen die sterven in een land op een bepaald moment. Dergelijke statistieken worden door de administratie verzameld om zicht te krijgen op belangrijke ontwikkelingen. De misdaadstatistieken kunnen aan deze lijst worden toegevoegd. Vanuit historisch perspectief zijn de statistieken van de veroordeelden de oudste gerechtelijke statistieken. De tweede toepassing van de statistiek is het onderwerp van deze inleiding: de statistiek is een methode om data te analyseren.

Statistiek is ook in de criminologie een methode om data die verzameld werden in het proces van een wetenschappelijk onderzoek, te bestuderen. Deze methode stelt onderzoekers in staat om logisch na te denken over data, en om één of twee dingen te doen: (1) te komen tot een beknopte synthese en betekenisvolle conclusies over de data (*beschrijvende statistiek*), en/of (2) karakteristieken van grote groepen bepalen -of afleiden-, gebaseerd op data afkomstig van kleinere

delen (*steekproeven*, ‘*samples*’) van de groep (*inferentiële statistiek*). We kunnen bijvoorbeeld data verzamelen bij alle inwoners van Gent in het kader van een onderzoeksproject waarbij we de attitudes van de Gentenaars ten opzichte van de politie, het stadsbeleid, enz. willen leren kennen. Deze resultaten zouden kunnen gebruikt worden voor een beschrijvende analyse om na te gaan in welke mate de burgers het eens zijn met de wijze waarop de politie de burger behandelt. Niet iedereen heeft dezelfde opinie. Het kan dus verwacht worden dat er grote verschillen opduiken, en mogelijks dat die verschillen in attitudes groter zijn bij jongeren dan bij ouderen. De resultaten kunnen echter ook grote tendensen laten zien. Wat denkt de modale Gentenaar? Wat is de gemiddelde score die de Gentenaar geeft als hij of zij de politie een “score op 10 moet geven”? Het zou natuurlijk heel duur zijn om iedereen te bevragen. Een alternatief is om een steekproef te trekken uit de Gentse bevolking. Data uit deze steekproef worden dan gebruikt om beweringen te doen over alle inwoners van Gent. Dit betekent dat er conclusies worden getrokken (*inferenties*), gebaseerd op informatie uit een kleinere steekproef, getrokken uit deze groep. Dat is onder bepaalde condities mogelijk, die we in dit handboek uiteenzetten.

Statistische analyse is een werkinstrument van wetenschappelijke ontdekking en wetenschappelijke kennisopbouw. Een voorbeeld van zo een wetenschappelijk proces is het gebruik van onderzoek om het waarheidsgehalte van een criminologische theorie na te gaan. Dit wil eenvoudigweg zeggen dat wij gaan controleren of de uitspraken die in een theorie vervat zitten over criminaliteit en diens oorzaken wel overeenstemmen met de bevindingen. Een theorie kan immers van alles beweren, maar we moeten vooral niet blind zijn en nagaan of een bewering wel klopt. Dat “nagaan” of iets wel klopt, doen we door gegevens te verzamelen. Dit vereist dat empirisch bewijs (data) dat gebaseerd is op onderzoekseenheden en dat die gegevens systematisch kunnen worden bestudeerd. Het gebruik van een aantal basisvaardigheden uit de wiskunde en algebra en statistische analyse in het bijzonder, maakt onderzoekers ertoe in staat om statistische vergelijkingen op te stellen en om nieuwe informatie te ontdekken die ervoor zorgt dat men zijn/haar onderzoekseenheid beter begrijpt.

In het wetenschappelijk proces is het doel meestal om iets te ontdekken dat voordien niet gekend was, of iets te bewijzen (waar of vals) waarvan voordien werd gedacht dat het waar was, maar dat nooit gestaafd werd met hard bewijs. De wijze waarop bewijs kan verworven worden is door

informatie (data) te verzamelen en deze te onderwerpen aan statistische analyse. Men mag echter niet uitgaan van wonderen. De statistiek toont verbanden aan, maar onderzoek is meer dan verbanden aantonen. Statistieken zeggen maar iets als de data die we gebruiken kwaliteitsvol zijn. Dat wordt soms wel eens vergeten. Statistiek komt pas te pas in criminologisch onderzoek nadat we een onderzoeksraag hebben bedacht en beslist hebben hoe we datgene dat we willen weten kunnen gaan meten.

#### **4. Theorieconstructie in een oogopslag**

Drie elementen zijn essentieel in gedegen criminologisch (statistisch) onderzoek: theoretische achtergrond, onderzoeksmethoden en kwaliteitsvolle statistische analyse. Hoewel deze elementen strikt gelinkt zijn met elkaar, is er een debat gaande -zelfs bij diegenen die het meeste achter het onderzoeksproces staan- over hun volgorde, belangrijkheid en wat precies van elk element in een boek moet opgenomen worden. Het is niet mogelijk om in dit handboek op alle elementen van wetenschappelijk kwantitatief onderzoek in te gaan. In dit basishandboek beperken we ons tot de univariate, bivariate beschrijvende statistiek, de principes van de inferentiële statistiek en een korte introductie tot de multivariate statistiek. De leerinhoud van wat in dit handboek centraal staat, moet gezien worden in het licht van de kwantitatieve criminologische methoden en technieken die in het tweede bachelorjaar van de opleiding tot bachelor in de criminologische wetenschappen gedoceerd wordt.

##### ***Wat is theorie?***

Op het basisniveau bestaat een **theorie** uit beweringen (proposities) over de relaties of associaties tussen *sociale fenomenen* zoals gebeurtenissen en eigenschappen van onderzoeksobjecten, vaak individuen, soms ook groepen, buurten, steden, landen,... In de criminologie bijvoorbeeld, zijn er theorieën die stellen via welke processen individuen criminaliteit als alternatief zien en besluiten in een bepaalde omstandigheid een delict te plegen. Daarnaast zijn er theorieën over de factoren die maken dat individuen de overtreding van normen als moreel acceptabel gaan zien: dat is een kwestie van het samenspel tussen de ontwikkelingscontext (omgeving, gezin, vrienden), de sociale bindingen die men heeft, en de persoonlijke psychologische, biologische en genetische kwetsbaarheden. Theorieën bevatten stelsels van uitspraken.

We geven een heel eenvoudig voorbeeld, gebaseerd op een observatie, die we willen verklaren (het “explanandum”), een aanvangsconditie, i.e. een conditie die we waargenomen hebben bij de persoon die we observeren. Theorieën worden grondig gecontroleerd door te kijken in welke mate deze observatie door een algemene wetmatigheid kan verklaard worden. Een sociale wetmatigheid in zijn meest banale vorm is een stabiel patroon, niets meer, niets minder. Het is een beproefde samenhang. Men mag dat niet fatalistisch interpreteren.

**Sociale wetmatigheid:** Wanneer jongeren zwakkere morele standaarden hebben, zijn zij sneller geneigd om criminaliteit als alternatief te zien.

**Aanvangsconditie:** Jan Janssens vindt het moreel niet verkeerd om eigendom van anderen mee te nemen. Mensen moeten maar op hun spullen letten. Eigen schuld dikke bult.

---

### **Explanandum: Jan Janssens heeft zopas de iPad van een medeleerling gestolen**

Het doel van deze beweringen is om verklaringen te ontwikkelen waarom dingen zijn zoals ze zijn (de processen van worden en zijn-ontwikkeling) en om via inzicht in processen een diepere causale verklaring te bekomen. Zonder theorie bestaan er vaak alleen vermoedens en verhaaltjes (“just so stories”). Met een criminologische theorie kunnen we beweringen of ideeën ontwikkelen die gebaseerd zijn op gedegen observatie en kunnen we misschien ook iets doen aan het probleem van de criminaliteit.

#### **Theorie en onderzoek**

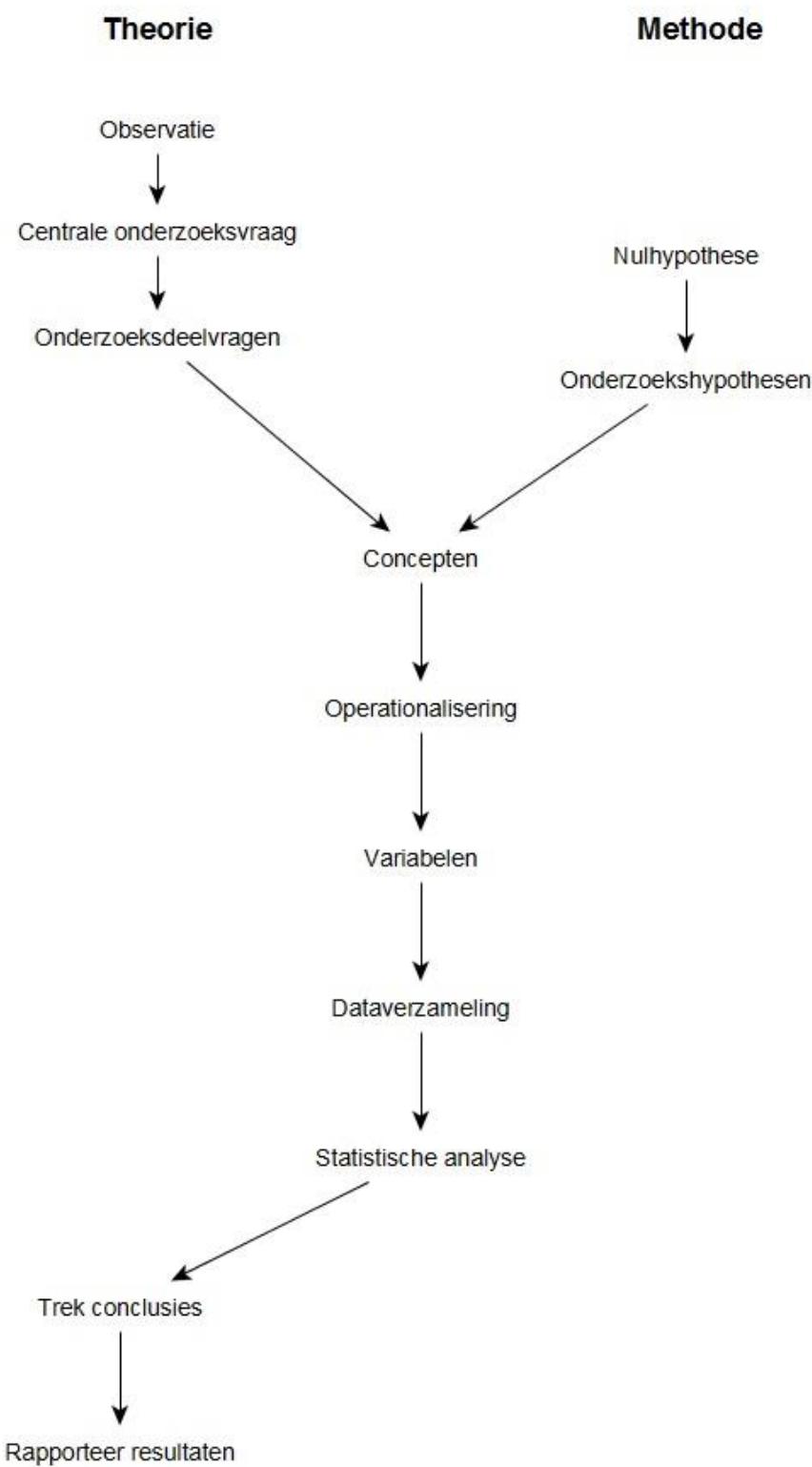
Een criminologische theorie kan op verschillende wijzen ontwikkeld worden. Ten eerste kunnen onderzoekers naar de wereld rondom hen kijken, een sociaal fenomeen nemen dat hun interesse wekt, en beweringen formuleren over waarom fenomenen op een bepaalde wijze werken. Dit wordt **inductie** genoemd. Bijvoorbeeld, een onderzoeker die criminaliteitstrends voor een aantal jaar in een stad volgt. Hij/zij kan misschien zien dat de criminaliteit een patroon volgt in de stad: de criminaliteit kan concentraties kennen in bepaalde wijken, de uitgaansbuurten, de hoerenbuurt, de verpauperde arbeiderswijken rond het centrum,... **Inductief bezig zijn is gevvaarlijk.** We kijken te veel naar wat we willen zien en we veronderstellen dat we als we honderd keer hetzelfde hebben geobserveerd, dat de 101<sup>ste</sup> keer ook zo observeren. Dat is absoluut geen garantie. Het feit dat de

zon al elke dag van je leven opging, kan geen garantie bieden dat ze morgen opgaat... Dat is niet bedoeld om pessimistische over te komen en je dag te vergallen, dat is een realiteit. We kunnen het niet bewijzen. We zijn soms blind voor onze eigen fouten en foutieve veronderstellingen. Daarom bestaat er een andere manier van onderzoek doen: **deductie**.

Het proces van deductie begint met een veronderstelling over hoe het nu in elkaar zit. In ons voorbeeld over de patronen van criminaliteit in de stad, kunnen we vertrekken van de Broken Windows-hypothese: een gebroken ruit of overvolle vuilnisemmers zenden signalen uit (priming): mensen zijn onbewust geneigd hun eigen afval bij de hoop te gooien die daar toch al ligt. Zo kan een buurt makkelijk afglijden als dat niet aangepakt wordt, of als de buurt niet weerbaar is. Als we het te ver laten komen, dan verhuizen de mensen die er genoeg van gehad hebben, en dan verkrot een buurt. Dat is een proces dat kan gestopt worden. De statistische samenhang tussen overlast en criminaliteit is een vrij stabiele samenhang. Maar even belangrijk als de statistiek, het hulpmiddel, is (1) de gezonde redenering, het logisch denkvermogen van de criminoloog die kritisch denkt, zich geen oor laat aannaaien bij het lezen van een tabel waar statistieken in staan, en (2) het beschikken over kwaliteitsvolle gegevens. In de praktijk is een criminologisch onderzoek een combinatie van inductie en deductie,

## 5. Het proces van wetenschappelijk onderzoek

Het proces van wetenschappelijk onderzoek (met gebruik van deductie) wordt getoond in **Figuur 1-1**. Zoals te zien is in dit diagram, is de theorie het startpunt van het proces. Theorie wordt gedreven door observaties en leidt onderzoekers tot het initiëren van het onderzoeksproces door het formuleren van een centrale **onderzoeksvraag** en onderzoeksdeelvragen. Het is uit dit proces van theorievorming dat onderzoekers komen tot het proces gaande van de ontwikkeling van een nulhypothese tot het communiceren van resultaten.



Figuur 1-1 Onderzoeksproces: Theorie, onderzoeksmethoden en statistische analyse

### ***Observatie en nieuwsgierigheid***

De eerste stappen in het proces van wetenschappelijk onderzoek worden vaak over het hoofd gezien, terwijl ze wel belangrijk zijn, nl. **observatie en nieuwsgierigheid**. Zo zijn vele onderzoeksprojecten nooit uitgevoerd omdat onderzoekers het initiële onderwerp plots niet langer interessant vonden, waardoor een onderzoek een andere richting is uitgegaan. Een criminoloog zou het idee kunnen gehad hebben om het fenomeen van de jeugdcriminaliteit te bestuderen vanuit het perspectief van de economische deprivatie, maar na het lezen van vele onderzoeksrapporten toch maar besluiten om een andere richting in te slaan. Zo iets gebeurt, zeker wanneer eerdere onderzoeken aantonen dat er geen eenduidig verband bestaat. Wat drijft onderzoekers? Kennis? Honger naar kennis? Of een honger naar het onbekende? Kennis is steeds feilbaar (net als het menselijke redeneervermogen), maar dat wil niet zeggen dat alles zo maar even relatief of waar is. Wie dat zegt, is een even grote charlatan als diegene die zegt de waarheid in petto te hebben. Het is vaak de theorie die observatie en wetenschappelijk onderzoek stimuleert. Als je onderzoeksmaateriaal doorneemt waarin je geïnteresseerd bent, kan het zijn dat je misschien denkt een betere manier te kennen om hetzelfde onderzoeksprobleem te benaderen. Bijvoorbeeld: denk terug aan het probleem van de economische deprivatie en criminaliteit. Het verband is vaak zwak, maar het is mogelijk dat het verband wel bestaat voor bepaalde vormen van criminaliteit of dat het verband enkel in bepaalde omstandigheden of voor een beperkt aantal personen geldig is: het verband zou bijvoorbeeld afwezig kunnen zijn op jonge leeftijd, maar het verband zou sterker kunnen zijn in de volwassenheid. Uiteraard moet je al heel wat kennis hebben over theorieën vooral eer je dergelijke denkoefeningen kan maken. Statistische analyses doe je dus altijd best gewapend met een goede kennis van de criminologie, dit wil zeggen de bestaande theorie en de kritieken erop, de stand van zaken met betrekking tot het bestaande onderzoek. Het gebruik van een gestructureerd wetenschappelijk proces om jouw criminologisch relevante observaties te evalueren en criminologische verklaringen te formuleren is de basis van goede theorieontwikkeling.

Een voorbeeld van inductieve theorievorming kan getoond worden in Robert Burgess' zonale theorie. Studenten aan de universiteit van Chicago maakten kaarten van Chicago waarop de verschillende karakteristieken van buurten werden vermeld zoals welzijn, kindersterfte en huisvesting. Burgess observeerde deze kaarten en zag dat deze een zeer gelijkaardig patroon

volgden doorheen de stad. Zijn observaties leidden tot de ontwikkeling van een theorie over hoe steden groeien en veranderen. Deze theorie stelt dat steden groeien in concentrische ringen zoals de ringen die ontstaan wanneer je een steen in het water gooit. In deze configuratie zal de ring het dichtste tot het centrum van de stad gekenmerkt worden door de meeste verloedering, het hoogste level van kindersterfte en andere sociale ziektes in vergelijking met de ringen erbuiten. Dit alles is bedacht door simpelweg het bestuderen van kaarten die gemaakt werden door studenten en door het gebruik van inductieve theorievorming.

### ***Centrale onderzoeks vragen***

De **centrale onderzoeks vragen** is de drijvende gedachte achter een onderzoeksproject. Ze moet de reden voor de statistische criminologische studie weergeven. Centrale onderzoeks vragen zijn belangrijk omdat criminologisch onderzoek vaak geëvalueerd wordt op basis van hoe goed de onderzoeker erin is geslaagd de centrale onderzoeks vragen te beantwoorden. De centrale onderzoeks vragen moet voorzichtig worden geformuleerd zodat deze exact de focus van de studie weergeeft. Bijvoorbeeld in onderzoek naar het gebruik van dodelijk geweld door politie zou een mogelijke centrale onderzoeks vragen de volgende kunnen zijn:

“Welke persoonlijkheidskenmerken, attitudes en omgevingskenmerken tijdens een arrestatie zijn het meest van invloed op het gebruik van dodelijk politiegeweld tijdens een arrestatie?” Deze vraag is breed en ietwat vaag, maar deze kan perfect het doel van het onderzoeksproject weergeven.

### ***Onderzoeksdeelvragen***

Vaak is het zo dat de centrale onderzoeks vragen theoretisch, vaag en bijna onmogelijk direct te beantwoorden is. **Onderzoeksdeelvragen** delen de centrale onderzoeks vragen in subproblemen die gemakkelijker te behandelen zijn en zorgen ervoor dat de centrale onderzoeks vragen toetsbaar wordt met behulp van onderzoek. Als de centrale onderzoeks vragen het doel van het onderzoek omvat, dan suggereren de onderzoeksdeelvragen hoe dit doel kan verwezenlijkt worden.

Een onderzoeker die geïnteresseerd is in sociale en psychologische determinanten van jeugdcriminaliteit kan de volgende vragen stellen

- Hoe veel delicten rapporteren jongeren op jaarbasis?

- Hoe groot is het empathisch vermogen van jongeren?
- Hoeveel jongeren behoren tot een straatgroepje
- Is er een verband tussen het behoren tot een straatgroepje en de hoeveelheid delicten die men rapporteert?
- Wat is de relatie tussen de empathische vermogens van de jongere en de frequentie waarmee hij of zij geweld pleegt?
- Wat is het verband tussen het behoren tot een straatgroepje en het empathisch vermogen?

Deze onderzoeksdeelvragen delen de centrale onderzoeksvergadering in kleinere delen die makkelijker kunnen worden onderzocht. De antwoorden op deze onderzoeksdeelvragen zijn afgeleid van het onderzoeksproces en statistische analyse en stellen de onderzoeker in staat om de centrale onderzoeksvergadering te beantwoorden.

*'Doing research is like defusing a bomb. When you begin, you are all excited and focused on the end result. If you have a good plan, know the layout, and work the plan, you will typically get the results you sought. If you run in and start cutting without a plan, it is likely to blow up in your face.'*

## 6. Onderzoek: bewegen van theorie naar data en terug

Theorie kan niet op zich bestaan, net zoals onderzoek of statistiek. Theorie zonder onderzoek en statistische analyse is niet meer dan een fabel. Onderzoek zonder theorie is zoals het bouwen van een huis zonder plannen, en onderzoek zonder statistische analyse is zoals het bouwen van een huis zonder nagels: het is mogelijk en het is al gebeurd, maar het zou effectiever zijn met. Statistische analyse zonder theorie en onderzoeksmethodologie als gids voor onderzoek is dan weer zoals het hebben van intercontinentale ballistische raketten: leuk om te hebben en het zou je buren kunnen imponeren, maar niet echt bruikbaar. Wat de techniek ook is voor het ontwikkelen van theorie, **onderzoek** is de methode om een theorie te toetsen en valideren. In zijn puurste vorm is onderzoek een wetenschappelijke, systematische studie om nieuwe informatie te ontdekken of om de validiteit van theorieën die eerder werden ontwikkeld, te toetsen. Het voornaamste doel van onderzoek is ontdekken. Afhankelijk van het gebruik van een inductief of deductief proces, is onderzoek een systematische manier om observatie en statistische analyse om te zetten in theorie (inductie) of het toetsen van theorie met behulp van statistische analyse (deductie). Hoewel er geen exacte stappen

bestaan die moeten gevolgd worden bij het doen van onderzoek, zijn er wel een aantal algemene richtlijnen die zouden gevolgd moeten worden om zeker te zijn dat er niks ontbreekt in de studie. Deze stappen worden in dit hoofdstuk besproken en geven weer waar onderzoek en statistiek in het volledige proces van wetenschappelijk onderzoek passen.

### ***Hypothesen formuleren***

Eens de onderzoeksdeelvragen werden ontwikkeld, moet er beslist worden wat het onderzoek probeert te bepalen. Hypothesen zijn vragen of verklaringen waarvan de antwoorden de theoretische proposities van het onderzoek ondersteunen of weerleggen. Hypothesen zijn in het algemeen onder te verdelen in onderzoekshypothesen en nulhypothesen. Een **onderzoekshypothese** is een propositie, een veronderstelling die je wilt toetsen. Bij het gebruik van onderzoekshypothesen wordt de relatief abstracte bewoording van theorieontwikkeling omgezet in een meer concrete en toetsbare vorm, geschikt voor statistische analyse.

Eén van de vaak moeilijk begrijpbare, maar essentiële elementen van statistische analyse en hypothesetoetsing is dat onderzoek alleen niets kan bewijzen. Ook als onderzoekers een groot stuk van bevestiging vinden voor een associatie tussen twee variabelen, kan het zijn dat deze resultaten worden gevonden omdat er informatie ontbreekt of omdat het model wat gebrekkig is. Mogelijks kunnen andere onderzoekers de bevindingen weerleggen door bijkomstig onderzoek te doen. Als statistisch onderzoek niets definitiefs kan bewijzen, waarvoor kan het dan wel gebruikt worden? Dit werd gevonden met behulp van een **nulhypothese**, dat in het algemeen de vorm van één van de volgende voorbeelden aanneemt:

- Er is geen **statistisch significant** tussen twee kenmerken in een steekproef.
- De verschillen tussen de groepen bestaan door toevallige fouten.

Het doel van een nulhypothese is om een verband te weerleggen. Weerleggen dat er geen relatie is (het weerleggen van de nulhypothese), helpt het ondersteunen van een conclusie dat er een relatie is tussen de bestudeerde fenomenen.

### ***Constructie van het onderzoeksdesign***

Eens een beslissing is gemaakt van wat precies bestudeerd zal worden, kan het plannen van het actuele onderzoek beginnen. Als onderzoeker dien je voorzichtig te zijn en niet te snel te springen van deze stap naar andere stappen in het onderzoeksproces. Zo start je niet aan het bouwen van een huis zonder eerst andere huizen te bekijken en te bedenken hoe jouw huis er uit moet zien. Dus, waarom zou je een onderzoeksproject starten zonder een degelijke overweging van wat je wilt doen en vinden? Activiteiten in deze stap zijn het bepalen van onder andere de onderzoeksmethode (experiment, survey of een andere methode) en hoe het onderzoek in het algemeen moet benaderd worden. Als de onderzoeker data moet verzamelen, moeten beslissingen gemaakt worden aangaande hoe deze data dienen verzameld te worden, van welke groep deze data moeten komen en andere parameters. De beslissingen die hier worden gemaakt sturen de rest van het project, dus ze moeten zeer zorgvuldig worden gemaakt. Deze stap in het onderzoeksproces is ook opgelegd door het type van verzamelde data, dat op zich ook de statistische analyses bepaalt die gebruikt zullen worden.

### ***Conceptualisering***

Eens de onderzoeksdeelvragen en hypothesen zijn geformuleerd, moeten ze onderverdeeld worden in meer handelbare delen. Dit wordt gedaan door concepten uit de vragen en hypothesen te halen. **Concepten** zijn termen waarover in het algemeen een consensus bestaat en kan betrekking hebben op een kenmerk, fenomeen of een groep van onderling gerelateerde fenomenen. Concepten kunnen heel abstract (delinquentietolerantie) zijn of net concreet (man-vrouw).

### ***Operationalisering***

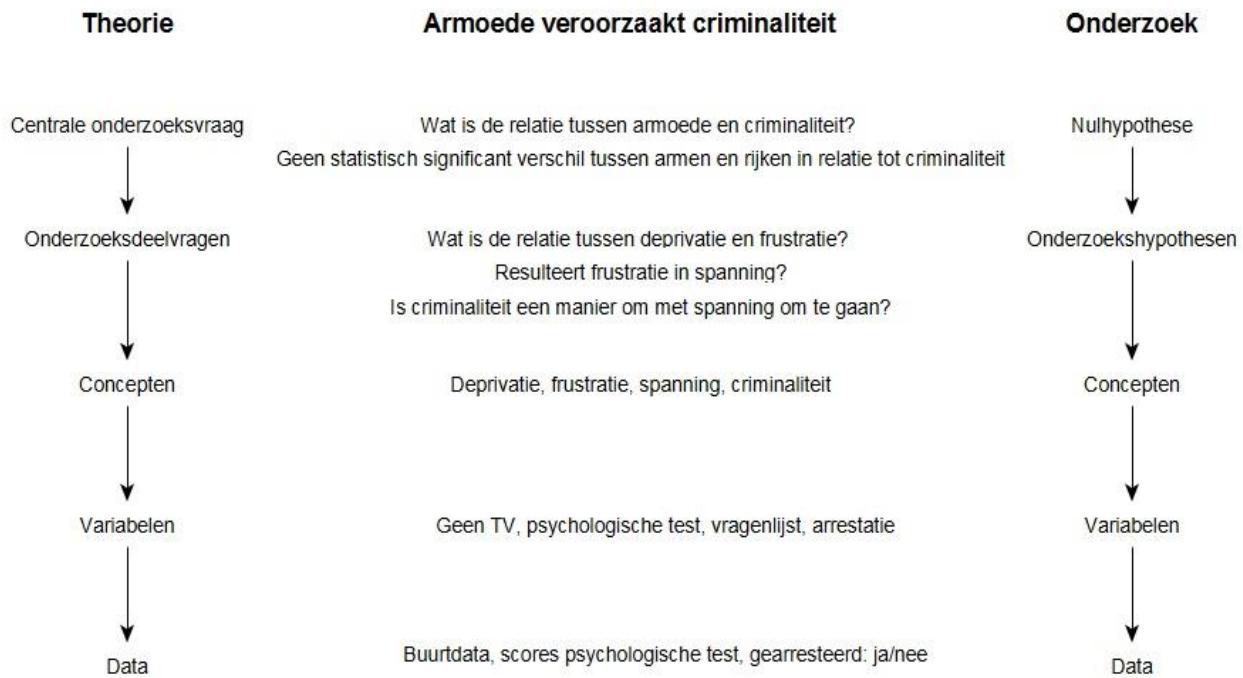
Om concepten te gebruiken in statistische analyse, moeten ze zo geformuleerd worden dat ze mathematisch kunnen worden geanalyseerd. Dit gebeurt door middel van **operationalisering**. Dit is het proces van het vertalen van een concept, dat abstract en woordelijk is, in een variabele die kan gezien en getoetst worden, door het omschrijven hoe het concept kan gemeten worden. Het proces van transformeren van concepten in variabelen demonstreert een kritiek punt in operationalisering: de geoperationaliseerde definities die gebruikt worden in onderzoek, zijn de onderzoeker zijn/haar definities en deze hoeven niet overeen te stemmen met de definities die anderen gebruiken of definities die dezelfde onderzoeker gebruikt in ander onderzoek.

Bijvoorbeeld, in dit onderzoeksproject is het mogelijk dat een politieman/-vrouw geoperationaliseerd werd als een inspecteur van politie binnen de lokale politie die belast is met interventie. Anderen zouden politieman/-vrouw anders kunnen definiëren en er bijvoorbeeld ook rechercheurs aan toevoegen en/of anderen met een politiefunctie.

### ***Data verzamelen***

Dataverzameling is de stap in het onderzoeksproces waar de meeste mensen eigenlijk willen starten en eigenlijk de laatste is waarmee zou moeten begonnen worden. Wanneer we terugkeren naar het voorbeeld van het bouwen van een huis, zou het starten met het verzamelen van data te vergelijken zijn met het beslissen om een huis te bouwen, 5 zakken cement en een ton nagels te bestellen en aan de slag gaan, zonder te werk te gaan met een degelijk ontwikkeld plan. Het kan zijn dat je effectief tot een huis komt, zeker wanneer je een expert bent, maar het zou beter zijn wanneer je begint met het zorgvuldig opmaken van een plan. Op dit punt zouden alle beslissingen betreffende het onderzoek moeten gemaakt zijn. De onderzoeker zou een weloverwogen theoretisch model moeten hebben evenals een duidelijk en compleet onderzoeksdesign waarin staat hoe de data zullen verzameld en geanalyseerd worden. De concepten zouden al geoperationaliseerd moeten zijn in variabelen die accuraat gemeten kunnen worden. Het enige wat nog moet gedaan worden is data verzamelen volgens het onderzoeksdesign. Deze analyseren volgens het design en de resultaten rapporteren.

Het onderzoeksproces betreffende het gebruik van dodelijk geweld door politie wordt weergegeven in **figuur 1-2**. Het toont elke stap in een deductief proces. De onderzoeker heeft een onderzoeksplan opgesteld volgens de stappen in het onderzoeksproces gaande van theorie naar data en van analyse naar de publicatie van conclusies. Dit wetenschappelijk proces is zowel bruikbaar in academisch als praktisch onderzoek. Let wel dat tussen de stappen ‘concepten’ en ‘het trekken van conclusies’ deze stappen liggen die direct van invloed zijn op de statistische analyse. Het is dus van het grootste belang dat deze stappen zorgvuldig doorlopen zijn aangezien ze bepalend zijn voor de statistische analyses.



**Figuur 1-2 Wetenschappelijk onderzoeksproces: voorbeeld van theorie naar data**

Figuur 1-2 geeft een illustratie van het proces van wetenschappelijk onderzoek. Hoewel niet alle stappen zijn weergegeven, toont de figuur wel de verschillende stappen dat men moet doornemen en het verschil tussen theorie en onderzoek. In de figuur kunnen ook de types van werkproducten afgelezen worden die na elke stap worden bekomen. In dit voorbeeld wordt vertrokken vanuit de vraag of het armoede criminaliteit veroorzaakt. De onderzoeksdeelvragen en hypothesen zetten de abstracte centrale onderzoeksraag om in toetsbare stellingen. De concepten delen de onderzoeksdeelvragen en hypothesen verder op in sleutelelementen die gemeten moeten worden. De variabelen die de operationalisering van de concepten betreffen, zijn die elementen waarover data verzameld moeten worden. Tenslotte kunnen de verzamelde data (de aantallen en andere informatie) worden bestudeerd door middel van statistische analyse.

### **Conclusies trekken**

Velen denken dat het proces van statistische analyse, en zelfs wetenschappelijk onderzoek, stopt aan het einde van de analyse. Niets is minder waar. Statistici en onderzoekers onderscheiden zich in de interpretatie van analyses en de conclusies die getrokken kunnen worden. Dit is over het algemeen ook het moeilijkste deel van statistische analyse. Deze stap houdt in of de resultaten van

de statistische analyse de hypothesen die bij de start van het onderzoeksproces zijn ontwikkeld, worden ondersteund. In deze stap van het onderzoeksproces stoppen we met de statistische analyse en methodologische kwesties en wordt er teruggekeerd naar theorie. Als de onderzoeker gebruik maakt van een deductief proces, is dit het punt waarop de theorie die in de eerste stappen werd uitgelijnd, vergeleken wordt met de resultaten. Nu valt het verdict of de theorie ondersteund of weerlegd wordt. In een inductief proces is dit het punt waarop de onderzoeker de eerste conclusies trekt over wat hij of zij heeft gezien. Kennelijk verlaat de theorie nooit echt het onderzoeksproces, net zoals methodologische en statistische kwesties belangrijk zijn in elke stap, van het opzetten van een theorie tot het trekken van conclusies.

Een student veranderen in een statistisch geschoold sociaal wetenschapper vereist het verscherpen van vaardigheden met betrekking tot het interpreteren van analyses. In staat zijn om problemen uit te werken of de computer ertoe te brengen om een antwoord te bieden op een specifieke analyse is één ding; het uitvoeren van die analyse, het terugkoppelen in het proces van wetenschappelijk onderzoek en het interpreteren van wat je hebt gevonden op zo een manier dat het nieuwe inzichten brengt in het topic dat wordt bestudeerd, is een andere zaak.

### ***Communiceren van resultaten***

De laatste stap in het proces van wetenschappelijk onderzoek is het communiceren van de resultaten van het onderzoek. Deze stap wordt ook vaak over het hoofd gezien. Velen geloven dat de bevindingen van het onderzoek niet gecommuniceerd moeten worden omdat ze het niet waard zijn, tenzij de bevindingen de hypotheses ondersteunen of een gigantische ontdekking inhouden. Hoewel het inderdaad zo is dat veel van de meest prestigieuze tijdschriften terugschrikken van het publiceren van negatieve resultaten, betekent het niet dat deze niet gecommuniceerd moeten worden. Het is belangrijk voor de praktiserende criminoloog in het werkveld dat zelfs negatieve resultaten worden gecommuniceerd. Dit moet mensen behoeden dezelfde fouten opnieuw te maken en zou moeten vermijden dat geld verspild wordt door iets te onderzoeken dat al eens eerder is onderzocht. Deze finale en essentiële stap van het onderzoeksproces kan op verschillende manieren worden verwezenlijkt. Het meest wenselijke is om de resultaten van een statistische studie te publiceren in een academisch tijdschrift of boek. De resultaten kunnen ook gecommuniceerd worden in meer praktisch gerichte publicaties (vb. laagdrempelige publicaties voor politie), in

paper presentaties op professionele conferenties. Echter, ook een masterproef kan gezien worden als een eerste stap in deze richting. Daartoe zijn de vaardigheden die hier getraind zullen worden essentieel. Het komt er immers op neer wanneer de resultaten van een statistische analyse uitgerekend zijn, om deze ook nog helder neer te schrijven zodat de informatie een didactische waarde krijgt. Alleen oefening baart kunst.

## 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 onderzoekt 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 en 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 regelovertraving voorkomen, en welke de achtergrondkenmerken zijn van de regelovertravers.

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 bevraagd 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 onderzoeks vragen**

Criminologische problemen roepen vragen op bij onderzoekers. Om deze specifieke vragen te kunnen beantwoorden, dient een probleemstelling (centrale onderzoeks vrag 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 onderzoeks vragen **beschrijvend, verkennend of verklarend** zijn. Een beschrijvende onderzoeks vrag 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 (falsificering). 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 onderzoeks vragen hebben betrekking op het aftasten van (nieuwe) veronderstellingen. Men zou kunnen zeggen dat de verkennende onderzoeks vraging zich ergens situeert tussen de beschrijvende en de verklarende. **De verklarende onderzoeks vraging** is er op gericht geobserveerde verschillen of veranderingen in uitkomst variabelen 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 onderzoeks vraging. 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 mogelijks in verband met bepaalde andere verschijnselen, zonder een verklaring te geven voor het fenomeen. Gedegen wetenschappers leiden hun onderzoeks vragen 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 probleem stelling, i.e. de contouren van een probleem, en zijn vaak minder op de hoogte van een criminologische deelterrein waarop ze een onderzoek zouden willen uitvoeren. Zo zijn veel studenten geïnteresseerd in herstelgerichte justitie, maar zij weten vaak niet erg goed welke specifieke onderzoeks vragen 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 onderzoeks vragen 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 onderzoeks vragen 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 criminale 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 criminologietema'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 **continuum** 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 *twoe 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 (v.b. 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 **exacte betekenis**. Een bekend 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;

...

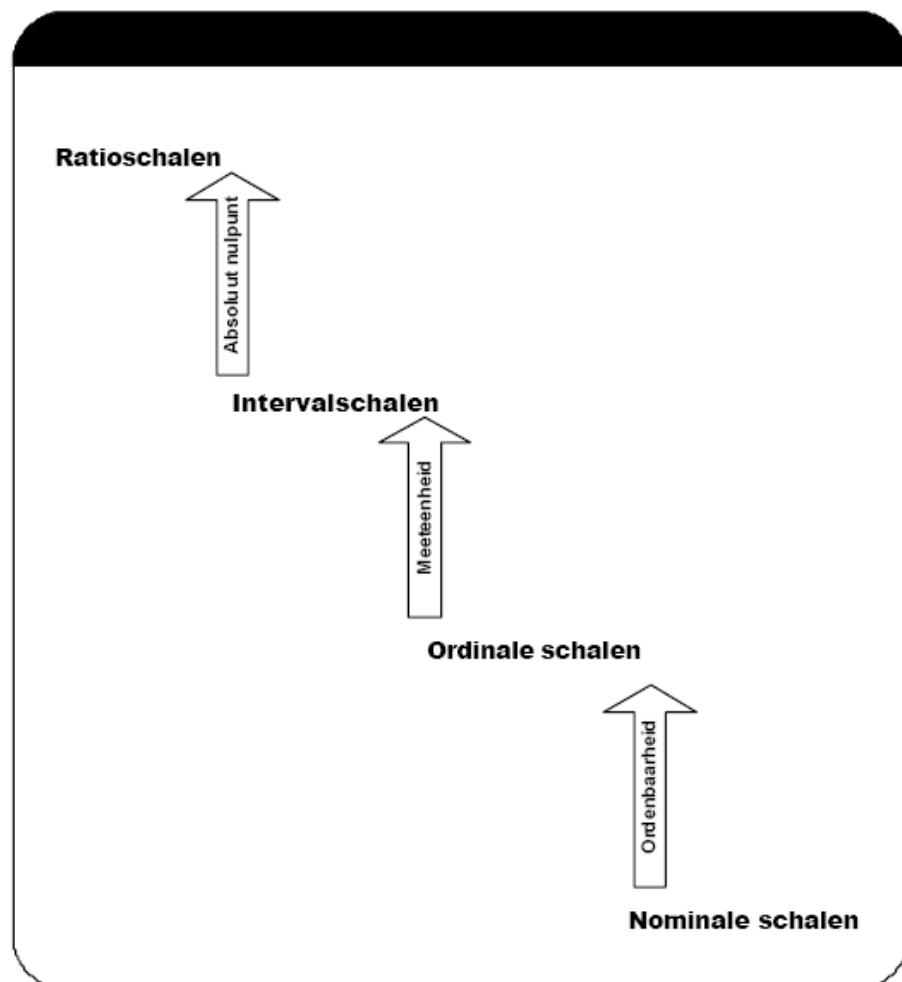
De afstanden tussen de waarden zijn in dit voorbeeld steeds even groot. De intervallen hebben daarmee een betekenis. Je kunt zeggen dat het verschil tussen klasse 3 en klasse 4 net zo groot is als het verschil tussen klasse 2 en 3. En er is geen absoluut of natuurlijk nulpunt. Je kunt niet zeggen dat 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’, ‘middelmatig’, ‘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 verwarring, 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
<b>Nominaal</b>	Ja	<i>Neen</i>	<i>Neen</i>	<i>Neen</i>
<b>Ordinaal</b>	Ja	Ja	<i>Neen</i>	<i>Neen</i>
<b>Interval</b>	Ja	Ja	Ja	<i>Neen</i>
<b>Ratio</b>	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\*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:

<b>Resnr</b>	<b>geslacht</b>	<b>opleiding</b>	<b>diploma vader</b>	<b>#bedreigd</b>	<b>leeftijd</b>
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:

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

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.

## Hoofdstuk 3

### De univariate beschrijvende statistiek

#### 1. Inleiding

Een datamatrix voor statistische analyse bestaat uit één of meerdere metingen, scores of waarden voor alle verschillende onderzoekseenheden (individuen, objecten, gebieden, criminale gebeurtenissen,...). De methodologie voor het organiseren en samenvatten of beschrijven van de gegevens voor een steekproef of de gehele populatie, wordt beschrijvende statistiek genoemd. Welke statistische technieken men hierbij gebruikt, is afhankelijk van **twee zaken**: (1) de **onderzoeks vraag** en (2) het **meetniveau** van de kenmerken van de eenheden waarover we uitspraken willen doen. Eenheden verschillen van elkaar in termen van bepaalde kenmerken. Studenten verschillen bijvoorbeeld in scores op hun examens, in leeftijd, geslacht,... Men kan deze verschillende waarden bestuderen en dit doet men door **frequentieverdelingen** en **grafieken** te maken die deze gegevens samenvatten en visualiseren. Door gebruik te maken van de frequentieverdelingen worden twee zaken duidelijk: de grote **centrale tendensen** en **bijzondere observaties**, i.e. observaties die weinig voorkomen. We leren dus nogal wat uit wat ogenschijnlijk maar een eenvoudige frequentieverdeling is.

Verder kiest men voor parameters van **centraliteit**, om de centrale tendensen weer te geven en parameters van **spreiding** om de waargenomen verschillen samen te vatten. Bij elk van de mogelijkheden moet telkens eerst de vraag gesteld worden wat het meetniveau is van de variabele. Immers, het meetniveau bepaalt welke mogelijkheden er bestaan. Vergelijk het geheel van statistische analysetechnieken met een gereedschapskist waaruit je diverse soorten van gereedschap kan kiezen. Je kiest wat je nodig hebt, maar daarin word je beperkt door de mogelijkheden van de variabelen. Hierbij geldt de algemene regel dat de mogelijkheden die we hebben bij het laagste meetniveau ook voor de hogere meetniveaus mogen gebruikt worden. Het omgekeerde is echter niet mogelijk.

#### 2. Over absolute en relatieve frequenties en hun grafische voorstelling

Kenmerken van onderzoekseenheden kunnen verschillende waarden hebben. Bij sommige kenmerken zijn de waarden al een getal, bij andere kenmerken zou je voor de voorkomende categorieën een getal kunnen verzinnen. De waarden van bijvoorbeeld het kenmerk leeftijd zijn getallen die direct gerelateerd zijn aan de werkelijkheid. Als een persoon 21 jaar oud is, is het logisch dat deze persoon de waarde 21 krijgt voor de variabele ‘leeftijd in jaren’. Geslacht heeft

geen vaststaande numerieke waarde. Om in de statistiek toch op een geordende wijze iets te kunnen zeggen over de onderzoekseenheden, krijgen de categorieën ‘man’ en ‘vrouw’, waarin de variabele ‘geslacht’ kan worden onderverdeeld, wel een numerieke waarde om de dataverwerking te vergemakkelijken. Je zou kunnen besluiten vrouwen de waarde 1 te geven en mannen de waarde 2. Op die manier kun je alle onderzoekseenheden voorzien van een numerieke waarde voor het kenmerk ‘geslacht’.

Een variabele (kenmerk) met de daarbij behorende waarden kun je op een overzichtelijke manier presenteren in een **frequentietabel**. Stel, in de zomervakantie zit je met wat vrienden op een terrasje en jij bent aangewezen om de drankjes te halen. Je vraagt je vrienden wat ze willen drinken. Je kunt proberen alles te onthouden, maar op een bierviltje de drankjes turven is gemakkelijker. Door te turven maak je een overzicht van het aantal keer dat een waarde voorkomt. Het tellen van de streepjes brengt je op de **absolute frequentie** van het kenmerk “te consumeren drankje” bij de onderzoekseenheden “vrienden”. Jan en Piet drinken een pint, Johan, Katrien en Mohammed drinken een koffie en Jana drinkt een thee. Die variabele die we hier beschreven hebben is “drankje”. Dit is een variabele van het nominale niveau. Hoezo nominaal niveau? Je kan toch ordenen, bijvoorbeeld op basis van het alcoholpercentage? Dat is juist, maar dan zou de variabele “alcoholpercentage” zijn en de eenheid het drankje. Hier zijn de eenheden jouw vrienden en het drankje is gewoon wat ze besteld hebben op het terras. Dit kan allemaal heel banaal klinken, we zouden het hier niet neerschrijven mochten studenten hier niet zo frequent fouten tegen maken. Je bent sneller verstrooid dan je denkt. Je leest een vraag maar half en je denkt het antwoord al te weten. Zo werkt het niet. Statistiek heeft te maken met nauwkeurigheid en dat begint vanaf het begin, van de meest banale analyse tot de meest complexe. Het moet juist zijn.

In kwantitatief criminologisch onderzoek is de situatie analoog als de situatie die we daarnet hebben beschreven. Je stelt in een criminologisch onderzoek geen gewone alledaagse vraag, maar een onderzoeksraag die gerelateerd is aan het fenomeen dat je wil bestuderen. Je bent criminoloog en je wilt weten hoeveel verschillende delicten jongeren binnen de tijdspanne van een jaar plegen. Je stelt dus een **beschrijvende onderzoeksraag**: hoeveel verschillende delicten heeft u het afgelopen jaar gepleegd?

Hoeveel jongeren rapporteren als misdrijf omschreven feiten, of meer algemeen ‘hoeveel elementen van de steekproef hebben een bepaalde waarde op de variabele criminaliteit?’ is een elementaire onderzoeksraag in jeugdcriminologisch onderzoek van beschrijvende aard. Het

aantal elementen met een bepaalde waarde van een variabele noemt men **de absolute frequentie van die waarde**. Men kan deze bepalen door een frequentietabel op te stellen. Dit kan voor elk meetniveau.

We komen terug op het voorbeeld van het delinquent gedrag dat jongeren plegen. Het is een voorbeeld uit een masterproef in de criminologische wetenschappen. Een oud-student deed onderzoek naar jeugdcriminaliteit en vroeg in verschillende scholen hoe vaak de jongeren bepaalde delicten gepleegd hadden in een bepaalde periode (twaalf maanden). De vraag uit de vragenlijst luidde: "hoe vaak heb je het afgelopen jaar iets gestolen": de antwoordcategorieën waren 0 = nul keer, 1= één keer,... , 6 = zes tot tien keer en 7 = meer dan tien keer. De ruwe antwoorden van de jongeren waren:

0 0 1 3 0 0 4 1 "geen antwoord" ...

Om een overzicht te krijgen over al deze waarnemingen, plaatsen we ze in een tabel waarin in de eerste kolom de verschillende waarden worden genoteerd die voorkomen. In een tweede kolom tellen we hoeveel keren elke categorie voorkomt. In de oude dagen dienden onderzoekers deze antwoorden te turven, nu gebeurt dit aan de hand van software. Hieronder zie je een tabel zoals deze gemaakt wordt met het statistische verwerkingspakket SPSS. Het voorbeeld is gebaseerd op de variabele diefstal. Let op de verschillende waarden die je in de rijen ziet staan.

**Variabele: Hoe vaak heb je het afgelopen jaar iets gestolen?**

Waarde		Frequentie Absolute aantallen	Percentages Relatieve aantallen	Geldige Percentages	Cumulatieve geldige percentages
Valid	nul keer	2848	92,4	94,4	94,4
	één keer	60	1,9	2,0	96,4
	twee keer	33	1,1	1,1	97,5
	drie keer	21	,7	,7	98,2
	vier keer	10	,3	,3	98,5
	vijf keer	10	,3	,3	98,9
	zes tot tien keer	11	,4	,4	99,2
	meer dan tien keer	23	,7	,8	100,0
	Totaal	3016	97,9	100,0	
	Missing System	66	2,1		
<b>n (steekproefgrootte)</b>		3082	100,0		

In de eerste kolom zie je de waarden die de vraag “hoe vaak heb je iets gestolen” in ons voorbeeld heeft. SPSS maakt een onderscheid tussen geldige waarden (valide) en ontbrekende waarden (missing). In de tweede kolom (Frequentie) staan de absolute frequenties, het aantal keer dat een bepaalde waarde voorkomt. De waarde “nul keer” komt klaarblijkelijk het meest voor. 2848 respondenten hebben een score nul keer. Dat wil zeggen dat 2848 respondenten het afgelopen jaar niets gestolen hebben.

Daarnaast staan de **percentages**. Percentages gaan van nul procent tot honderd procent. Het percentage dat een bepaalde waarde voorkomt is de verhouding tussen het absolute aantal keer dat een waarde voorkomt, gedeeld door het *steekproeffrequentie* (*n*), vermenigvuldigd met honderd. Als je aandachtig kijkt naar deze tabel, dan zie je dat 2,1 procent deze vraag niet heeft ingevuld of niet beantwoord heeft. 2,1 procent is het percentage ontbrekende informatie. Hier valt dit nog mee. De vraag is niet zo bedreigend dat een meerderheid er niet op geantwoord heeft. In ander onderzoek gebeurt het dat 20% de vraag niet beantwoordt. Je kan je dan afvragen of dat geen invloed heeft op de resultaten. Voorzichtigheid is dus zeker geboden. In ons voorbeeld hebben 66 respondenten de vraag niet ingevuld. In het valide percentage worden deze niet meegeteld.

De som van alle **absolute frequenties** is gelijk aan het totaal aantal elementen in de steekproef en wordt voorgesteld door de kleine letter “*n*”. Soms wordt de hoofdletter *N* gebruikt. De hoofdletter wordt gebruikt als het gaat om een populatie, de kleine letter wordt gebruikt als het gaat om een steekproef. De steekproefgrootte bedraagt 3082 respondenten. De statistische notatie is de volgende:

$$n = f_1 + f_2 + \dots + f_m = \sum_{i=1}^m f_i = n$$

Hier staat:  $2848+60+33+\dots+66 = 3082$ .

We kunnen ook de **relatieve frequentie** berekenen door **percentages** (kolom 3) te presenteren. Deelt men elke absolute frequentie door het totaal aantal waarnemingen (hier 3082) dan bekomt men **proporties**. Deze zijn hier niet weergegeven.

$$f_i' = \frac{f_i}{n}$$

Vermenigvuldigt men deze proporties met de waarde honderd dan spreken we over **percentages**. We geven dit als volgt weer:

$$f_i' = \frac{f_i}{n} \times 100$$

In de kolom **geldige percentages** zien we dat de respondenten die niet op de vraag hebben geantwoord, niet meegeteld. Voor de rest is de berekening dezelfde. Het steekproeffrequentie ( $n$ ) is dus iets kleiner. Vaak houdt men bij het maken van een tabel ook rekening met de mensen die ‘geen antwoord’ gaven, de zogeheten *weigeraars*. Men herberekent de percentages enkel voor diegenen die een antwoord hebben gegeven. Bij deze berekeningen wordt het aantal mensen die niet antwoordden of niet moesten antwoorden, afgetrokken van het totaal aantal waarnemingen. Dit nieuwe totaal (de geldige percentages) is dan het getal waardoor elke absolute frequentie wordt gedeeld. Het aantal **geldige percentages** vinden we in de derde kolom terug. Dit zijn de percentages berekend op diegenen die een *geldig antwoord* hebben op de vraag.

In de kolom daarnaast zien we de **cumulatieve percentages**: hier worden de percentages van elke volgende waarde bij de voorgaande opgeteld. Een criminoloog die statistieken analyseert laat zich niet misleiden door wat in één kolom staat, maar brengt in rekening wat in elke kolom staat.

Het voordeel van het gebruik van relatieve frequenties en percentages is dat de frequentieverdelingen voor verschillende, niet even grote groepen personen beter vergelijkbaar worden. Wanneer **minstens op een ordinaal niveau** werd gemeten, heeft het verder ook zin om de **cumulatieve percentages** weer te geven. Anders gezegd: *het heeft absoluut geen zin om voor nominale kenmerken cumulatieve percentages te berekenen*. Aan de hand van deze cumulatieve verdeling kunnen we zien hoeveel waarnemingen kleiner dan of gelijk aan een bepaalde waarde zijn. We stellen dit als volgt voor:

$$K(x_i) = \sum_{x_j \leq x_i} F_j$$

Zo kunnen we onmiddellijk aflezen dat meer dan 90% van de bevraagde jongeren het afgelopen jaar geen enkele diefstal heeft gepleegd. We onderscheiden *absolute cumulatieve frequenties* (voorgesteld door de hoofdletter **K**) en *relatieve cumulatieve frequenties* of cumulatieve percentages (voorgesteld door de kleine letter **k**).

Opgelet: **percentages zijn niet altijd geschikt om gegevens voor te stellen. Bevat de totale steekproef minder dan dertig eenheden, dan werkt men beter met absolute aantallen.** De reden hiervoor is eenvoudig: je kan misleidende resultaten presenteren. Als één respondent op drie antwoordt dat deze al cocaïne gesnoven heeft, betekent dat 33.33% van de steekproef. Dat percentage is misleidend en totaal betekenisloos aangezien je maar drie respondenten hebt. Deze fout wordt heel vaak gemaakt door criminologen die kwalitatief onderzoek doen, en die denken dat ze hun onderzoek wat wetenschappelijker kunnen maken door een statistiekje te maken op basis van hun beperkte steekproeven.

Wanneer gepercenteerde tabellen worden weergegeven in een rapport, moet absoluut het steekproeffrequentie (*n*) worden vermeld. Enkel dan kan men zelf de gegevens nog (beperkt) herordenen. In criminologisch onderzoek heeft men het vaak over **incidentie** en **prevalentie** van criminaliteit. Incidentie is het aantal nieuwe gevallen van een bepaalde conditie dat voorkomt in een populatie gedurende een bepaalde periode. Bijvoorbeeld het totaal aantal nieuwe gedetineerden in de Gentse gevangenis gedurende een jaar. Prevalentie is het totaal aantal personen in een bepaalde conditie in een populatie op een bepaald moment. Bijvoorbeeld het aantal gedetineerden in de Gentse gevangenis op 1 januari 2015.

### **Grafische voorstellingen**

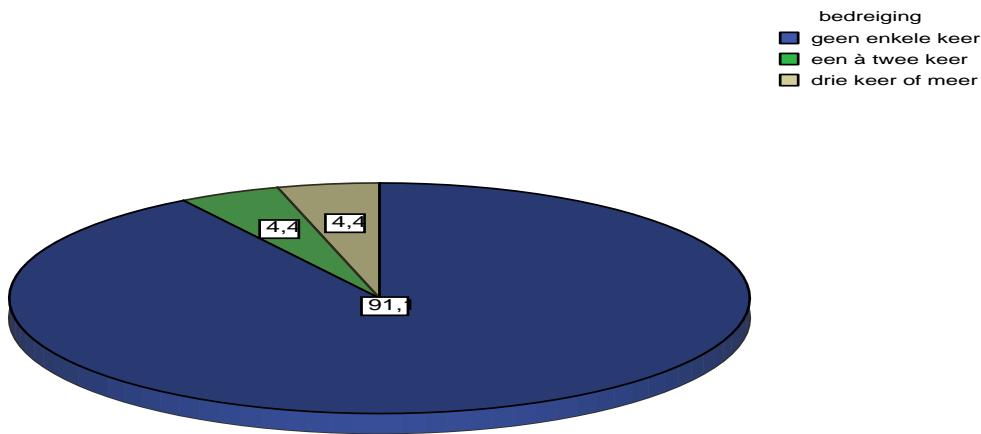
Een grafiek zegt soms meer dan duizend woorden. Dat is een heel terechte uitspraak. In een oogopslag zie je wat opvalt: de trends in criminaliteit, de stijging of daling. De grillige vormen of stabiele patronen inzake misdaad en de kenmerken waarmee misdaad samenhangt worden in een oogopslag zichtbaar gemaakt. Nog anders gezegd: met grafische voorstellingen weet je precies welk soort vlees je in de kuip hebt, of anders gezegd wat voor steekproef je hebt, in termen van de belangrijke variabelen die je bestudeert.

Een grafische voorstelling streeft altijd naar het overbrengen van de informatie van complexe gegevens via beelden (tekeningen). Hierbij is het belangrijk dat het geconstrueerde beeld in overeenstemming is met de reële informatie in de gegevens. Verder mag het informatiegehalte van een grafiek niet te klein zijn maar ook niet te groot. Bij het tekenen van grafieken is het belangrijk om ook rekening te houden met het **meetniveau** van de variabele. Grafische voorstellingen worden aan de hand van statistische verwerkingspakketten gemaakt.

### **Taartdiagram of cirkelgrafiek (pie chart)**

Bij een taartdiagram wordt op basis van de verschillende frequenties of percentages een cirkelschijf verdeeld in sectoren.

**Figuur: pie chart van de variabele “hoe vaak heb je het afgelopen jaar iemand bedreigd”**



Deze grafiek is gemakkelijk te interpreteren en visualiseert de informatie uit de tabel. **Taartdiagrammen** zijn populair bij kenmerken gemeten op het **nominale** en **ordinale** niveau. We gebruiken deze grafiek eigenlijk wanneer we met een beperkt aantal categorieën werken. Dit is belangrijk met betrekking tot de duidelijkheid. Bij meer dan 5 categorieën wordt het moeilijk om de verschillende cirkelsectoren van elkaar te onderscheiden. De volgorde van de categorieën hangt af van het meetniveau. Bij nominale variabelen is de volgorde in principe willekeurig. Toch is het interessant om categorieën die inhoudelijk bij elkaar aansluiten naast elkaar te plaatsen. Voor ordinale en metrische variabelen volgt men het ordeningscriterium bij het weergeven van de categorieën.

### **Staafdiagram (bar chart)**

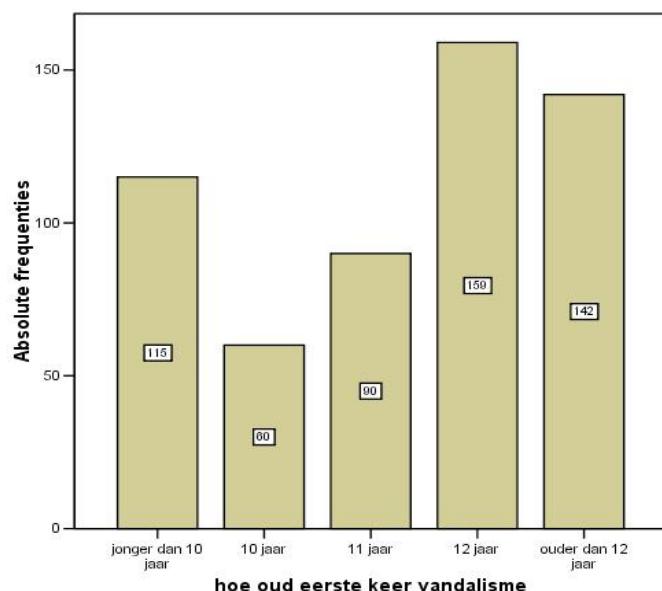
Bij een staafdiagram worden de gegevens op ***twee assen*** voorgesteld. Op de ***horizontale as of de X-as***, worden **de verschillende categorieën van de variabele X weergegeven**. De ***verticale as, of de Y-as*** geeft de aantallen weer. Dit kan onder de vorm van absolute frequenties of percentages. Op elke waarde van X tekent men een staafje. De oppervlakte van deze verschillende staven drukken de absolute aantallen of percentages uit. De staafjes kunnen zowel horizontaal als verticaal worden voorgesteld.

Als de staafjes driedimensionaal worden voorgesteld, spreken we van een ***blokgrafiek***.

Er zijn een aantal regels die in acht moeten worden genomen.

- Staafdiagrammen worden gebruikt bij de grafische voorstelling van kenmerken gemeten op het **nominale en ordinale** niveau.
- We plaatsen de staafjes los van elkaar omdat de variabele nominaal of ordinaal is.
- Indien het om een metrische variabele gaat, mogen de staafjes bij elkaar aansluiten en wordt hierdoor visueel weergegeven dat de categorieën zich op ***een continuüm*** bevinden (we spreken dan van een *histogram* – zie verder).
- In een staafdiagram kunnen meer categorieën worden voorgesteld.

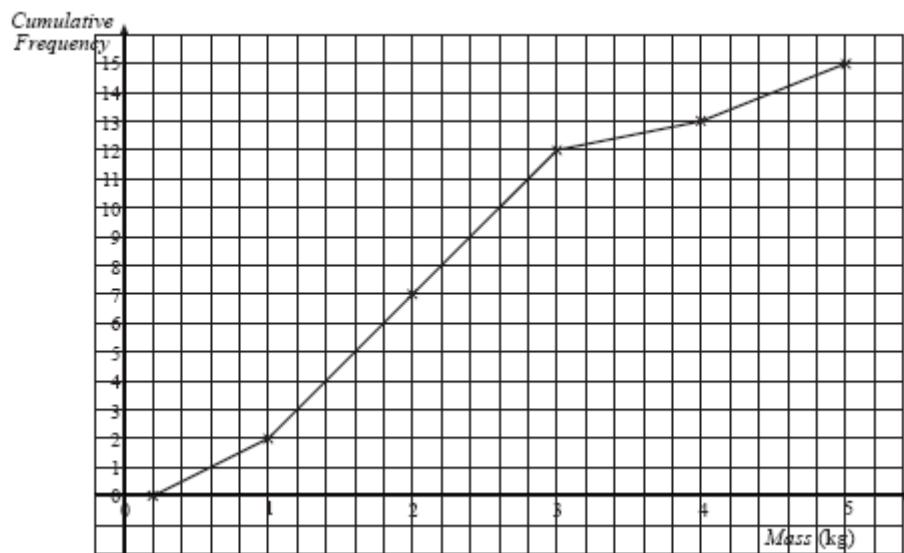
**Figuur: staafdiagram van het kenmerk “leeftijd bij het eerste delict”**



### **Cumulatief frequentiediagram**

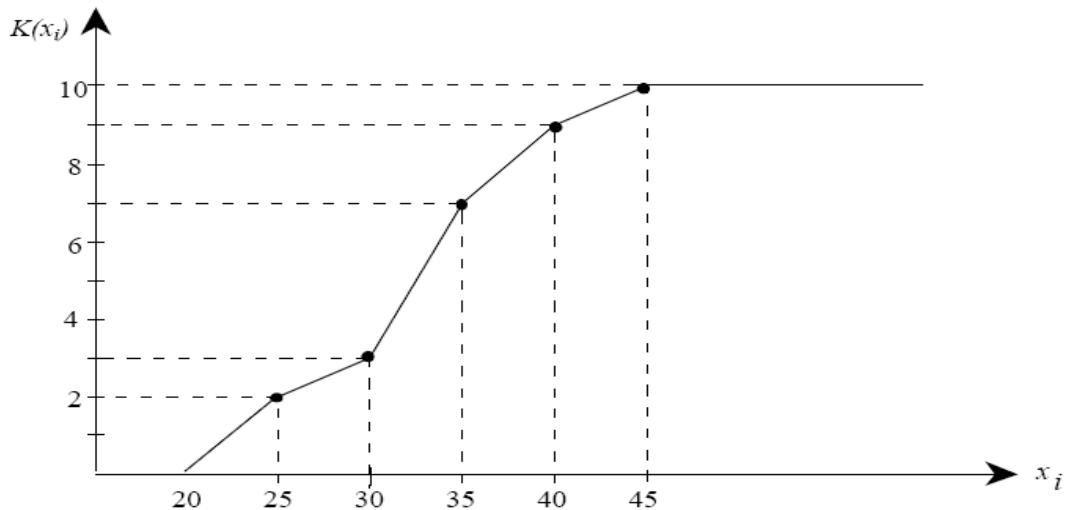
De absolute of relatieve cumulatieve frequenties kunnen worden voorgesteld in een cumulatief frequentiediagram en dit vanaf het **ordinale** niveau. Voor gegevens die niet in klassen werden ingedeeld, tonen we hieronder een voorbeeld.

**Figuur: cumulatief frequentiediagram**



We zien op dit cumulatieve diagram dat op de verticale as de cumulatieve frequenties staan en dat op de horizontale as de verschillende waarden staan. De variabele is hier massa (uitgedrukt in kg). Wanneer de waarden in klassen zijn ingedeeld, zal het cumulatieve frequentiediagram er enigszins anders uitzien. Bij indeling in klassen gaan we immers uit van de hypothese dat de waarnemingen gelijkmataig verdeeld zijn over de verschillende klassen. Het gevolg hiervan is dat het diagram niet langer een trapfunctie zal zijn, maar een gebroken lijn.

**Figuur: cumulatief frequentiediagram**

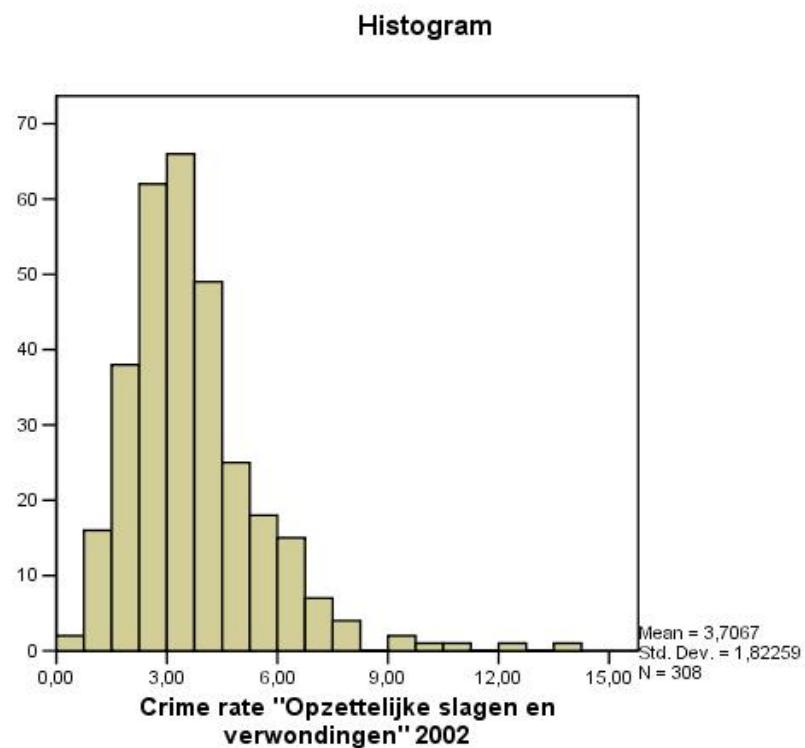


### Histogram

Het histogram is een veelgebruikte visuele voorstelling voor **metrische** kenmerken. **Metrische** gegevens in een klassentabel worden vaak voorgesteld d.m.v. een **histogram**: de breedte van elk balkje komt overeen met de klassenlengte; de hoogte van elk balkje met het aantal gegevens dat binnen die klasse valt. Bemerk dat de balkjes elkaar moeten raken! Mochten er spaties tussen de balkjes zitten, zouden mensen verkeerdelijk kunnen denken dat de bijbehorende lengtes niet voorkomen of niet mogelijk zijn.

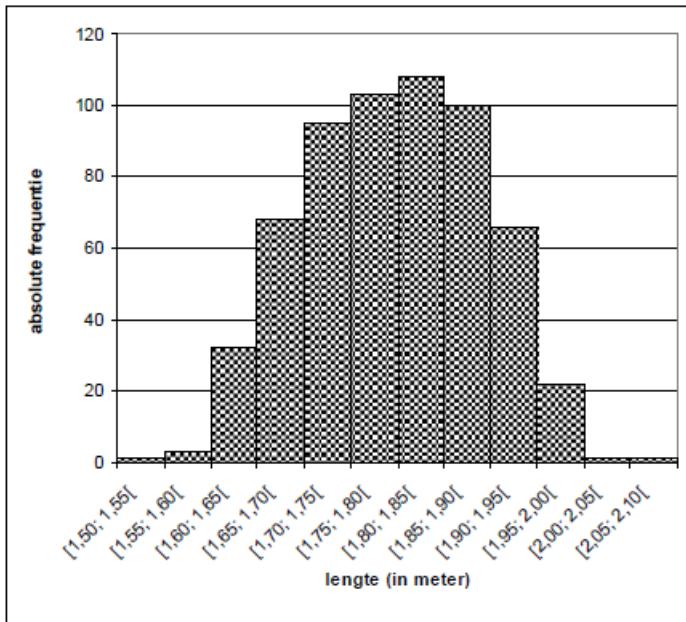
Dit is een belangrijk verschil met het staafdiagram. De blokjes worden aan elkaar getekend omdat de waarden van X elkaar opvolgen. Door deze voorstelling wordt duidelijk gemaakt dat de categorieën op een **continuüm** liggen. De oppervlakte van de kolom is steeds gelijk aan de frequentie van de waarde die de kolom voorstelt. De totale oppervlakte van het histogram is dan gelijk aan het totaal van het aantal elementen. De totale oppervlakte bevat honderd procent van de waarnemingen. Bij gegevens in klassen ingedeeld, worden de klassengrenzen weergegeven. Het voorbeeld dat we presenteren is het histogram van de crime rate voor opzettelijke slagen en verwondingen in Vlaamse gemeenten.

**Figuur: histogram voor de criminaliteitsgraad voor geweld**



**Figuur: histogram voor lengte (in klassen gegroepeerd)**

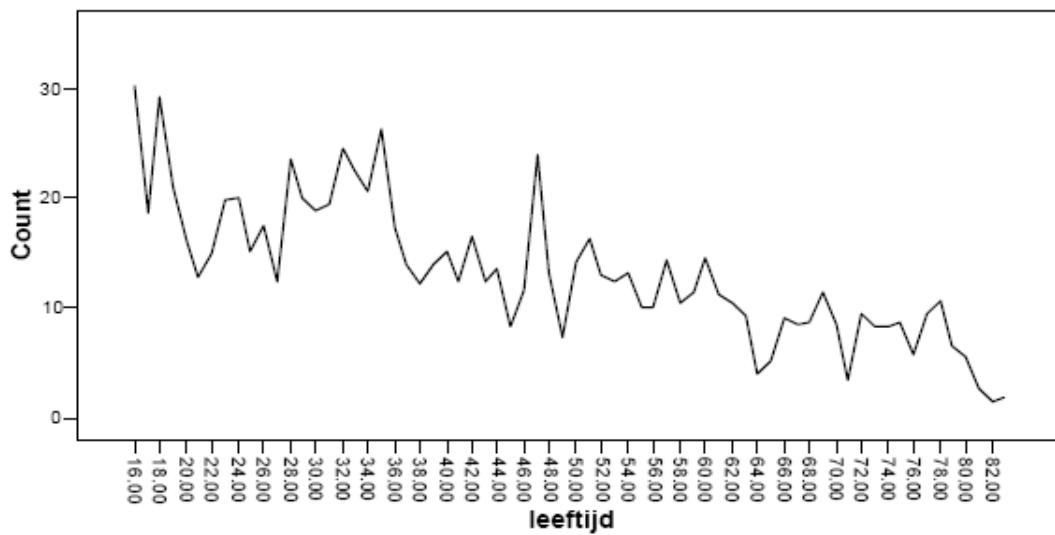
Lengte (in m)	aantal
[1,50; 1,55[	1
[1,55; 1,60[	3
[1,60; 1,65[	32
[1,65; 1,70[	68
[1,70; 1,75[	95
[1,75; 1,80[	103
[1,80; 1,85[	108
[1,85; 1,90[	100
[1,90; 1,95[	66
[1,95; 2,00[	22
[2,00; 2,05[	1
[2,05; 2,10[	1
<i>totaal</i>	600



### **Lijndiagram**

In een lijndiagram worden de niet in klassen gegroepeerde gegevens visueel voorgesteld door hun frequentie op een verticale as aan te duiden en de waarden op de x-as te stellen. De punten worden vervolgens door middel van een lijn met elkaar verbonden.

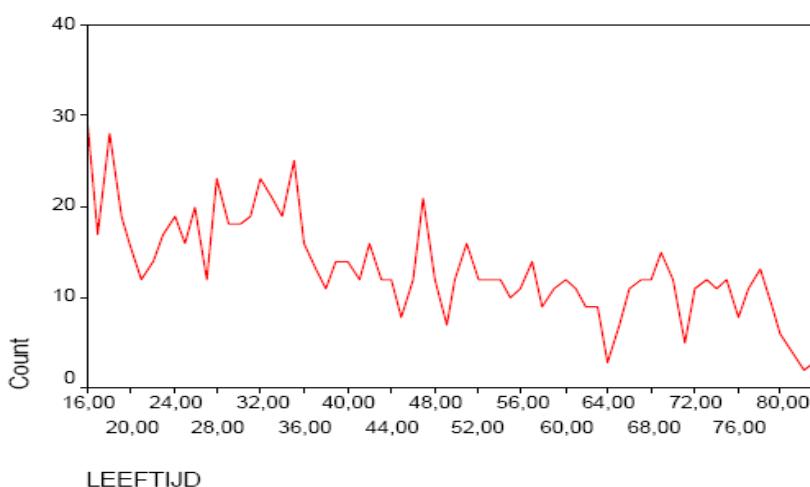
**Figuur: lijndiagram voor de variabele leeftijd**



### **Frequentiepolygoon**

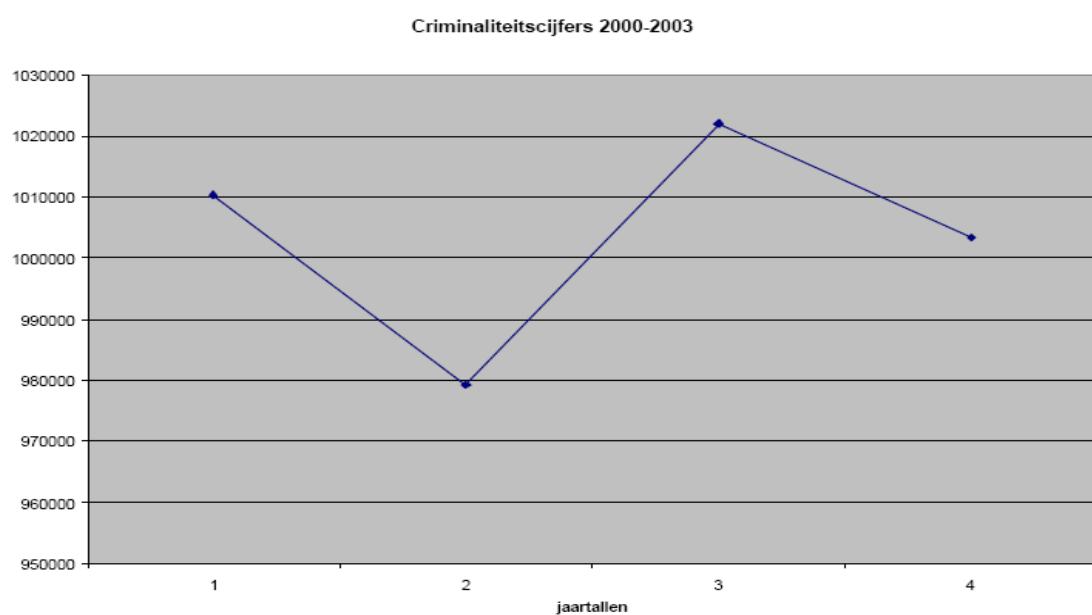
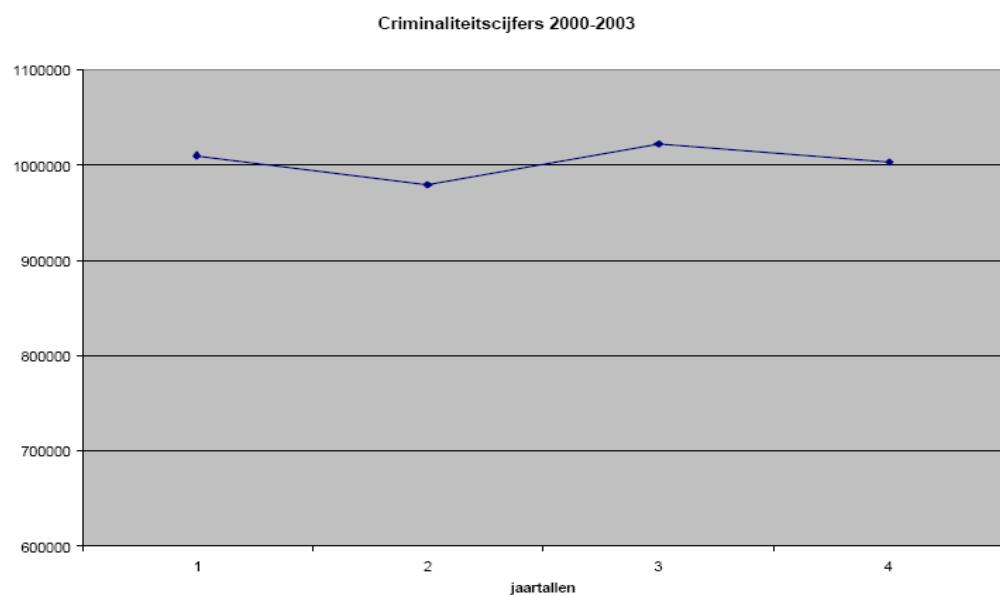
De frequentiepolygoon is nauw verwant aan het lijndiagram. Het is in feite een lijndiagram voor in klassen gegroepeerde gegevens. In een frequentiepolygoon worden de categorieën, voorgesteld door hun klassenmiddens en punten gevormd door het aanduiden van de hoogte van de frequentie, met elkaar verbonden.

**Figuur: frequentiepolygoon voor de variabele leeftijd**



### ***Opgelet met grafische voorstellingen***

Grafieken kunnen heel verhelderend werken, doch kunnen ook heel bedrieglijk zijn. Vergeet daarom nooit ook de cijfers zelf van dichterbij te bekijken. De wijze waarop men de x- en y-as ijkt, is immers bepalend voor de mate waarin men detaillering ziet in het verloop van de cijfers en dus ook voor de wijze waarop het beeld zal overkomen bij de lezer. Onderstaand voorbeeld illustreert dit. In de eerste grafiek lijken de criminaliteitscijfers niet echt sterk te schommelen, er lijkt sprake te zijn van een vrij stabiele trend, daar waar de tweede grafiek veel grotere schommelingen lijkt te vertonen. Beide zijn nochtans opgesteld met exact dezelfde cijfers. Alles heeft uiteraard te maken de manier waarop de **Y -as** geijkt is.



Bron: Federale politie, criminaliteitscijfers 2000-2003

### 3. Parameters van centraliteit

Parameters van centraliteit geven een antwoord op **beschrijvende onderzoeks vragen**. Deze onderzoeks vragen zijn er op gericht **de centrale tendensen** te ontdekken. Centrale tendensen vinden we door te kijken naar centrale waarden. We spreken ook van **centrummaten**. Een centrummaat hanteren we wanneer we de frequentieverdeling willen kenmerken aan de hand van een *centraal gelegen waarde*. Deze waarde wordt dan als een representatieve maat beschouwd die de volledige verdeling van de waarnemingen zo goed mogelijk karakteriseert. Het gebruik van een bepaalde maat hangt af van het meetniveau van de variabele.

#### **De modus**

**De modus** is een centrummaat en betreft de categorie van de variabele met de *frequentie die het vaakst voorkomt*. Aangezien de modus enkel verwijst naar de waarde met de hoogste frequentie is er geen enkel probleem om de modus te bepalen bij nominale variabelen, ordinale en metrische variabelen. De modus heeft echter een beperkte betekenis en verwijst enkel naar de meest voorkomende waarde. Het geeft de meest in het oog springende categorie aan. De modus is meestal heel stabiel en dit is haar sterke. Er mogen al heel wat verschuivingen optreden in de frequenties van de verschillende categorieën voordat de modus wijzigt. Het is mogelijk dat er twee of meer modi voorkomen wanneer er twee of meer categorieën voorkomen met een even hoge frequentie.

Laten we een voorbeeld geven. In een jongerenbevraging werd gepeild naar spijbelgedrag van vrienden van de respondenten. De frequentietabel wordt hieronder weergegeven.

**Hoeveel van je vrienden hebben ooit gespjbeld?**

		Frequentie	Percentage	Geldig	Percent	Cumulatief percentage
Valid	<b>geen enkele</b>	<b>1953</b>	<b>63,4</b>		<b>63,5</b>	63,5
	één vriend	924	30,0		30,0	93,5
	twee vrienden	144	4,7		4,7	98,1
	drie of meer vrienden	57	1,8		1,9	100,0
	Total	3078	99,9		100,0	
Missing	<b>System</b>	4	,1			
	Total	3082	100,0			

Hieruit blijkt dat de **meest voorkomende waarde “geen enkele”** is. 1953 jongeren of 63,5 procent geeft dit antwoord. De meeste jonge adolescenten geven te kennen dat ze geen vrienden

hebben die spijbelen. De categorie “geen enkele” is hiermee de modus. De modus is niet 1953, maar de categorie die met dit aantal overeenkomt!

### **De mediaan**

**De mediaan** van een statistische verdeling is het *midden van die verdeling*. De mediaan is een centrummaat die het punt in de frequentieverdeling aangeeft waaronder 50% van de gevallen en waarboven de andere 50% van de gevallen liggen. De frequentieverdeling wordt als het ware in twee gelijke stukken gedeeld. De mediaan vormt met andere woorden het *middelpunt van de verdeling*. De mediaan is de middelste van de (oneven aantal) waarden in de rangschikking naar grootte. Bij een even aantal waarden is de mediaan het gemiddelde van de beide middelste waarden. Om de mediaan te kunnen bepalen van een kenmerk moeten de categorieën in oplopende volgorde gerangschikt zijn. Dit betekent dat men minstens een **ordinaal meetniveau** moet hebben om de mediaan te mogen gebruiken. Strikt genomen is de definitie enkel van toepassing als het aantal elementen ( $n$ ) in de frequentieverdeling oneven is. In die situatie is er één waarde die in het midden ligt. Bij een even aantal waarnemingen zijn er in feite twee middelste waarden. Gaat het om dezelfde waarden dan is dit de mediaan. Gaat het om twee verschillende waarden dan kan men één van de waarden per toeval selecteren of indien het meetniveau het toelaat het gemiddelde van de twee berekenen.

Een voorbeeld wordt hieronder uitgewerkt voor ruwe gegevens die niet in tabellen verwerkt zijn. We vragen aan 13 respondenten ( $n = 13$ ) hoeveel keer ze slachtoffer werden van vandalisme aan hun fiets. We bekomen de volgende antwoorden:

5; 1; 3; 3; 4; 5; 1; 2; 3; 4; 4 ;5; 2;

**Alle waarnemingen worden eerst van laag naar hoog geordend:**

<b>Waarden</b>	1	1	2	2	3	3	3	4	4	4	5	5	5
<b>Respondenten</b>	1	2	3	4	5	6	7	8	9	10	11	12	13

$n = 13$  of oneven waardoor de mediaan gelijk wordt aan de waarde van de zevende respondent  $(n+1)/2$ . In dit voorbeeld is dit de waarde 3. Indien het hier om 12 eenheden zou gaan, dan zijn de 6de en 7de waarneming de middelste. Beide zijn 3 dus ook hier is de mediaan gelijk aan 3.

In de hierboven gepresenteerde tabel die spijbelgedrag van vrienden meet, is de mediaan gelijk aan de waarde “geen enkele”. **Hoe kan ik dat afleiden uit een frequentietabel?** Het is

eenvoudig. 3078 respondenten vulden een geldig antwoord in. De waarde die overeenkomt met de waarneming die zich situeert tussen de 1539<sup>ste</sup> en 1540<sup>ste</sup> waarneming ( $3078+1)/2$ ), is eveneens “**geen enkele**”. De helft van het aantal respondenten heeft geen enkele vriend die ooit gespijbeld heeft, de andere helft heeft geen enkele tot en met drie of meer vrienden die ooit gespijbeld hebben.

### **De kwantieren**

De mediaan is een speciaal geval van de maten die we **kwantieren** noemen. Men kan een geordende rij elementen niet alleen in twee gelijke groepen indelen (zoals we doen wanneer we de mediaan gebruiken) maar in principe in om het even welk aantal groepen met een gelijk aantal elementen. Net zoals bij de mediaan verwachten we dat de variabele gemeten is op het ordinale meetniveau. Zo kan men geordende gegevens in drie gelijke groepen indelen: het laagste derde, het middelste derde en het hoogste derde. De twee waarden van de variabele die gebruikt worden om de drie groepen af te bakenen, zijn T1 en T2 (eerste en tweede derde). Het is meer gebruikelijk om **kwartieren** te gebruiken. Hiervoor zijn drie waarden nodig: het eerste quartiel (Q1 of de waarde waaronder zich 25% van alle eenheden bevindt), het tweede quartiel (Q2 of de waarde waaronder 50% van alle eenheden valt of ook de mediaan) en het derde quartiel (Q3; de waarde waaronder zich 75% bevindt). Bij voldoende eenheden kan men ook *deciliën* (1<sup>ste</sup> deciel = 1<sup>ste</sup> 10 %, 2<sup>de</sup> deciel = 1<sup>ste</sup> 20 % etc.) of *percentielen* bepalen.

#### **Kwartieren van het aantal verschillende zelfgerapporteerde feiten**

<b>n</b>	<b>Valid (geldig)</b>	3015
	<b>Missing (ontbrekende data)</b>	67
<b>Kwartieren</b>	<b>Q1 (25%) of eerste quartiel</b>	0
	<b>Q2 (50%) of tweede quartiel</b>	1
	<b>Q3 (75%) of derde quartiel</b>	3

De mediaan is gelijk aan 1. De helft van de respondenten heeft dus 0 tot 1 delict op jaarbasis gepleegd. Het eerste quartiel is gelijk aan 0, het derde is gelijk aan 3. Dit betekent dat een kwart van de respondenten geen delict heeft gepleegd. Driekwart van de respondenten heeft geen enkel tot en met drie verschillende delicten (uit een reeks van negen) gepleegd.

### **Het rekenkundig gemiddelde**

**Het rekenkundig gemiddelde** is een centrummaat dat gebruikt wordt bij variabelen gemeten op het metrische niveau, dus bij interval en ratio-variabelen. Het rekenkundig gemiddelde van een kenmerk wordt verkregen door alle voorkomende waarden bij elkaar op te tellen en vervolgens het totaal te delen door het aantal respondenten. Als er  $n$  respondenten zijn, wordt hun rekenkundig gemiddelde gegeven door de formule:

Voor *individuele waarnemingen*:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Of in het geval van *absolute frequenties*:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^m f_i \times x_i$$

Of in het geval van *relatieve frequenties*:

$$\bar{x} = \sum_{i=1}^m p_i \times x_i$$

Aantal zelfgerapporteerde delicten uit een reeks van 9 delictvragen

Aantal verschillende misdrijven (uit een reeks van 9 vormen)	Frequentie	Procent	Geldig procent	Cumulatief procent
Geldig	0	1021	33,1	33,9
	1	708	23,0	23,5
	2	469	15,2	15,6
	3	284	9,2	9,4
	4	196	6,4	6,5
	5	146	4,7	4,8
	6	72	2,3	2,4
	7	49	1,6	1,6
	8	36	1,2	1,2
	9	34	1,1	1,1
Totaal geldig	3015	97,8	100,0	
Missing	System	67	2,2	
Totaal		3082	100,0	

In een zelfrapportagestudie werd gevraagd aan jonge adolescenten of ze een reeks van negen delicten gepleegd hadden. Het gemiddeld aantal gerapporteerde delicten bedraagt 1,78. Dit kan ook als volgt uit de frequentietabel worden afgeleid door de som te nemen van de producten van elke waarde met diens absolute frequentie:

$$[(1021*0) + (1*708) + (2*469) + (3*284) + (4*196) + (5*146) + (6*72) + (7*49) + (8*36) + (9*34)]/3015 = 1.78$$

Het **rekenkundig gemiddelde** heeft enkele belangrijke kenmerken:

- (1) Het rekenkundig gemiddelde wordt enkel voor metrische variabelen gebruikt;
- (2) het rekenkundig gemiddelde is het evenwichtspunt (of zwaartepunt) van de verdeling;
- (3) het rekenkundig gemiddelde is geen resistente (robuste) maat: dit betekent dat de waarde gevoelig is voor uitschieters of extreme waarden;<sup>2</sup>
- (4) de som van alle afwijkingen tegenover het rekenkundig gemiddelde is nul.

Dit laatste vergt enige uitleg. De afwijkingen tegenover het gemiddelde noemen we **deviationscores**.

Hieronder presenteren we deze eigenschap aan de hand van de formule en maken een kleine rekensom om dit te bewijzen.

$\sum_{i=1}^n (x_i - \bar{x}) = 0$	$x_i$	$x_i - \bar{x}$
	12	1,17
	14	3,17
	7	-3,83
	10	-0,83
	9	-1,83
	13	2,17
<b>Som</b>	<b>65</b>	<b>0,00</b>
<b><math>\bar{x}</math></b>		<b>10,83</b>

---

<sup>2</sup> Extreme waarden noemen we outliers en komen heel weinig voor. Door hun extreem karakter beïnvloeden ze wel mee de uitkomst.

### **Verantwoord kiezen tussen centrummaten**

Welke centrummaat zullen we kiezen bij de rapportage van gegevens? Zowel de modus, de mediaan als het rekenkundige gemiddelde zijn centrummaten. De vraag stelt zich welke het best gehanteerd kan worden. Stel dat we deze 3 centrummaten berekenen voor een zelfde verdeling en we bekomen 3 verschillende resultaten. Hoe moeten we met dergelijke bevindingen omgaan? Dit is bijvoorbeeld het geval voor het aantal zelfgerapporteerde delicten uit het hierboven vermelde onderzoek.

#### **Parameters van centraliteit: Aantal zelfgerapporteerde feiten**

N	Valid	3015
	Missing	67
Gemiddelde		1,7847
Mediaan		1,0000
Modus		,00

Belangrijk is te weten dat bij de berekening van het rekenkundige gemiddelde alle waarnemingen worden betrokken. Dit is voordelig, want het informatiegehalte is daarom zeer groot. Maar er is een nadeel: en dat nadeel is precies dat *extreme waarden* de resultaten van het gemiddelde beïnvloeden. De mediaan is minder gevoelig aan extreme waarden, aangezien deze maat afhankelijk is van waar de meeste frequenties zich bevinden. Het is belangrijk om ook eens de verdeling van een kenmerk waarin je geïnteresseerd bent vormelijk te gaan bestuderen. Bij *niet-symmetrische verdelingen* (en dat komt vaak voor in de criminologie: zie verder, wanneer we de parameters van vorm presenteren) wordt het gemiddelde sterk beïnvloed door extreme waarden: liggen deze rechts, dan schuift het gemiddelde mee op naar rechts, m.a.w: dan zal het rekenkundig gemiddelde groter zijn dan de mediaan. Bij niet-symmetrische verdelingen is het rekenkundige gemiddelde dus niet altijd een betrouwbare centrummaat. Ze is te groot wanneer de 'staart' van de verdeling rechts ligt en te klein als deze links ligt.

***De mediaan is vaak een meer betrouwbare centrummaat bij niet-symmetrische verdelingen.*** Het is van het grootste belang dit te beseffen, aangezien veel kenmerken die door criminologen bestudeerd worden een niet-symmetrisch karakter hebben. Zo zijn slachtofferschap en daderschap gemeten binnen een tijdspanne zeer ongelijk verdeeld. Een voor de criminologie typerend voorbeeld is de vaststelling dat een kleine groep van delictplegers heel actief is. Berekenen we het gemiddelde dan beïnvloeden deze extreme waarden, die we "*outliers*" of "*uitschieters*" noemen, de resultaten.

#### **4. Parameters van spreiding: vive la difference!**

Wat betekent spreiding? Vrijwel alle kenmerken in onze samenleving kennen spreiding. Een bepaald kenmerk heeft dan meerdere categorieën en na een bevraging van respondenten (of andere eenheden) blijken de respondenten niet in één en dezelfde categorie terecht te komen, maar (heel vaak ongelijk) verdeeld te zijn over de verschillende categorieën. M.a.w. verschillende mensen vallen in de verschillende categorieën van het kenmerk dat we bestuderen. Hoe meer de waarnemingen verspreid zijn over alle categorieën, hoe groter de spreiding. Parameters van spreiding bieden een antwoord op beschrijvende onderzoeks vragen die de ongelijke spreiding van criminologisch relevante fenomenen willen bestuderen. Hoe ongelijk is slachtofferschap verdeeld in de samenleving? Hoe sterk verschillen jonge adolescenten in hun betrokkenheid bij delinquent gedrag? Zo zijn er nog meer voorbeelden te bedenken.

In de criminologische theorie speelt de ongelijke verdeling van respondenten over de categorieën van een kenmerk een grote rol. Men stelt zich de vraag waaraan deze variatie tussen de onderzochte eenheden kan worden toegeschreven. Wanneer we de criminale carrières van de Belgen bestuderen, bemerken we heel grote verschillen. Een reeks van mensen komt in de georganiseerde misdaad terecht, de meesten echter niet. Hoe komt dat? We zullen ons afvragen hoe we deze grote verschillen kunnen verklaren. Opleiding, hard werken, een hoog IQ, bindingen met de samenleving, iets te verliezen hebben,... het zijn allemaal mogelijke verklaringen voor de variatie in de betrokkenheid bij crimineel gedrag die we trachten te meten. In de criminologie stellen we ook vast dat veel mensen zelden of nooit slachtoffer worden van bepaalde delicten, en een kleine groep herhaaldelijk slachtoffer wordt. We stellen vast dat in veel gemeenten de geregistreerde criminaliteit aan de lage kant is, en in een paar gemeenten eerder hoge concentraties kent. Deze variabiliteit interesseert de criminoloog. Waarom is het zo dat criminaliteit zulke ongelijke verdelingen kent? Al jaren proberen criminologen, met de hulp van theorieën hiervoor verklaringen te bieden en verklaringen te gaan toetsen.

Hieruit vloeit voort dat een variabele voldoende spreiding moet hebben (variatie) om ze te onderzoeken. Als iedereen hetzelfde gedrag vertoont (bv. iedereen pleegt met eenzelfde frequentie delicten) dan kan men geen statistisch onderzoek opzetten met de vraag hoe de frequentie van delictplegen te verklaren valt. Men beschouwt een variabele als verklaard wanneer de *spreiding* of *variatie* in die variabele kan toegeschreven worden aan een identificeerbare bron. Deze bron is vanuit statistisch oogpunt een andere variabele (bivariate

beschrijvende statistiek) of een set van andere variabelen (multivariate beschrijvende statistiek). Dit komt in latere hoofdstukken van deze cursus aan bod. **Afhankelijk van het meetniveau van een variabele** wordt een spreidingsmaat gekozen.

### **De variatieratio (VR)**

Op het **nominale meetniveau** treffen we **de spreidingsmaat ‘de variatieratio’** aan. Alle spreidingsmaten voor nominale schalen voldoen aan het principe dat de waarde groter wordt naarmate de heterogeniteit groter wordt. *De variatieratio is de proportie waarnemingen die niet tot de modale categorie behoort.* De variatieratio neemt de waarde van nul aan indien alle waarnemingen tot de modale categorie behoren. Een minder aangename eigenschap is het feit dat er geen vaste bovenlimiet is. Er is geen maximumwaarde. Dus is het moeilijk de spreiding te gaan interpreteren. De maximale waarde van de variatieratio benadert de waarde één wanneer iedere waarneming een verschillende waarde heeft. De variatieratio is dus een eenvoudige maat om te berekenen, maar is niet genormeerd en houdt enkel rekening met de proportie van waarnemingen die tot de modale klasse behoort.<sup>3</sup> In het voorbeeld werken we een situatie uit.

Slachtofferschap afgelopen vijf jaar	Percentage
Geen slachtoffer	75
Minstens één maal slachtoffer de afgelopen vijf jaar	25

We zien dat de modale categorie de categorie “niet” is. Daartoe behoort 75%. De variatieratio is dus 100% - 75% oftewel 25%. Of in procentages uitgedrukt:  $1 - 0.75 = 0.25$ . De spreiding is aan de lage kant want is ver van 1 verwijderd. Wat betekent dat nu inhoudelijk: mensen worden niet zo vaak slachtoffer, de waargenomen variatie is niet zo groot. Een minderheid wordt slachtoffer.

### **De index van diversiteit (ID)**

Op het **nominale meetniveau** treffen we ook de **index van diversiteit** aan. De index van diversiteit is een spreidingsmaat die net zoals de variatieratio is gebaseerd op de relatieve frequenties van de categorieën, maar het enige verschil is nu dat rekening wordt gehouden met de proportie van waarnemingen binnen iedere categorie of equivalentieklasse. De waarde van

---

<sup>3</sup> Normering betekent in deze situatie bij nominale kenmerken dat een vaste waarde wordt bereikt bij maximale spreiding en dat het aantal categorieën niet van invloed is op de spreidingsmaat.

deze parameter geeft een idee van de *mate van concentratie van de waarnemingen over de categorieën* van de variabele. Meer concreet betekent dit dat als dit cijfer nul bedraagt, alle waarnemingen dezelfde waarde hebben en er dus ook geen sprake kan zijn van spreiding van waarnemingen.

De ID wordt als volgt berekend:  $ID = 1 - (f'1*f'1 + f'2*f'2 + f'3*f'3 + \dots f'n*f'n)$ .

Oftewel: de waarde één minus de som van de gekwadrateerde proporties in elke categorie. Door te kwadrateren wordt aan de categorieën met een hogere frequentie een hoger gewicht toegekend. We geven een voorbeeld.

<b>Regio van herkomst van vreemdelingen in het Vlaamse gewest</b>	<b>Freq. (Absolute frequentie)</b>	<b>Prop. (Proportie)</b>	<b>Gekwadrateerde proportie (= prop * prop)</b>
Europa	155 098	0.55	0.30
Afrika	57 065	0.20	0.04
Azië	55 794	0.20	0.04
Oceanië	369	0.00	0.00
Amerika	8 020	0.03	0.00
Onbekend	4 741	0.02	0.00
Som	281 087	1	0.38

Wanneer we nu de spreidingsmaat uitrekenen, moeten we voor elke categorie eerst de proportie berekenen en dan het kwadraat van deze proportie.

Vervolgens rekenen we uit:  $ID = 1 - (0.30+0.04+0.04+0.00+0.00+0.00) = 0.62$ . De index van diversiteit is interessant omdat deze in termen van kansen kan geïnterpreteerd worden. De ID betekent: “**de kans dat twee willekeurig gekozen vreemdelingen in het Vlaamse Gewest uit een verschillende regio afkomstig zijn, bedraagt 0.62”**

### **De variatiebreedte**

Wanneer het meetniveau van een variabele ordinaal is, worden **ordinale spreidingsmaten** gebruikt. Een bekend voorbeeld is de **variatiebreedte**. De variatiebreedte is het verschil tussen de grootste en de kleinste waargenomen waarde.

$$V = \max_i x_i - \min_i x_i$$

Laat ons een voorbeeld geven. We hernemen de vraag uit ons onderzoek naar spijbelgedrag.

#### Hoeveel van je vrienden hebben ooit gespjbeld?

		Frequentie	Percentage	Geldig	Percent	Cumulatief percentage
<b>Valid</b>	<b>geen enkele</b>	1953	<b>63,4</b>	<b>63,5</b>	63,5	
	één vriend	924	30,0	30,0	93,5	
	twee vrienden	144	4,7	4,7	98,1	
	<b>drie of meer vrienden</b>	57	1,8	1,9	100,0	
	Total	3078	99,9	100,0		
<b>Missing</b>	<b>System</b>	4	,1			
	Total	3082	100,0			

De variatiebreedte kan berekend worden want de antwoordcategorieën zijn gemeten op het ordinale niveau. *De variatiebreedte gaat van “geen enkele” tot “drie of meer vrienden”.*

De variatiebreedte is een zeer rudimentaire parameter van spreiding. Ze geeft immers enkel aan over welke afstand de waarnemingen verspreid zijn. Het zou best kunnen dat de variatiebreedte groot is, en het gevolg is van het feit dat er een aantal respondenten in een hoge rang zitten. Wanneer de gegevens in klassen gegroepeerd zijn, wordt de variatiebreedte gedefinieerd als het verschil tussen de bovengrens van de hoogste klasse en de ondergrens van de laagste klasse. Sommige auteurs hanteren een andere definitie, namelijk als het verschil tussen de klassenmiddens van beide uiterste klassen.

#### *De interkwartiel-afstand (K3-K1)*

De **interkwartiel-afstand** kan ook gehanteerd worden vanaf het ordinale meetniveau. De interkwartielafstand is het verschil tussen het derde en eerste kwartiel. Aangezien het eerste kwartiel het punt aangeeft waaronder 25% van de verdeling ligt, en het derde kwartiel het punt waaronder 75% van de verdeling ligt, bevat de interkwartielafstand de helft van het totale aantal waarnemingen. Extreme waarden hebben geen invloed op de interkwartiel-afstand. De **interdeciel-afstand** is het verschil tussen het negende en eerste deciel.

#### *Spreidingsmaten op metrisch niveau*

Als we te maken hebben met kenmerken die tenminste op het intervalniveau worden gemeten, dan kunnen we de afwijkingen tegenover het rekenkundig gemiddelde berekenen. De **metrische spreidingsmaten** zijn dus allemaal op eenzelfde principe gebaseerd. Men bepaalt

eerst het rekenkundig gemiddelde op basis van alle observaties. Vervolgens willen we weten elke waarneming verschilt van het rekenkundig gemiddelde. De logica hierachter is heel eenvoudig. Observaties die heel ver verwijderd liggen van het rekenkundig gemiddelde zijn heel bijzonder of afwijkend (deviant). Het is wel handig om deze verschillen te kunnen uitdrukken in een getal. Op die manier weten we meteen hoe homogeen of heterogeen onze steekproef is. Het motto hier is dus: **vive la difference!** Leve de variabiliteit. Variatie is de bron van alle theorie. Er zijn verschillen tussen individuen in termen van criminaliteit die men zelf begaat, maar ook in termen van criminaliteit waarvan men zelf het slachtoffer wordt. Er zijn naast verschillen tussen individuen ook verschillen binnen individuen. Een individu is geen statisch organisme. Individuen ontwikkelen zich doorheen de levensloop en dus plegen zij in sommige perioden van hun leven meer regelovertredend gedrag dan in andere perioden. Ook dat is spreiding. We willen in de criminologie verklaringen bieden voor variatie. In feite is dat zoeken naar variatie eigen aan de wetenschap. Ook in andere wetenschappen, zoals de biologie, wil men variatie beschrijven. Darwin beschreef de variatie en ontwikkeling van soorten, waaronder de mens, en het is op basis van variaties dat Darwin zich vragen begon te stellen die hebben geleid tot zijn evolutietheorie. Variatie doet er dus wel degelijk toe. De spreidingsmaten die we gebruiken op het metrische niveau zijn allemaal gebaseerd op de afwijkingen tegenover het rekenkundig gemiddelde. Dat betekent dat zij ook beïnvloed worden door de eigenschappen van dat rekenkundig gemiddelde. Wat geldt voor het rekenkundig gemiddelde, geldt ook voor de spreidingsmaten die er op gebaseerd zijn: net zoals het gemiddelde zijn de metrische spreidingsmaten niet zo robuust. Ze zijn zeer afhankelijk van extreme waarden. Een paar extreme waarnemingen (statistische “**outliers**”), die voorkomen wanneer een kenmerk heel scheef verdeeld is in een steekproef, kunnen de spreiding sterk beïnvloeden. Laat criminaliteit nu zo een variabele zijn die heel scheef verdeeld is. Eén van de belangrijkste variabelen uit de criminologische theorievorming is enorm scheef verdeeld. Dat heeft consequenties, maar daar komen we ten gepaste tijd nog op terug. We bespreken achtereenvolgens de *gemiddelde absolute afwijking, de variatie, de (steekproef)variantie, de (steekproef)standaardafwijking en de variatiecoëfficiënt*.<sup>4</sup>

---

<sup>4</sup> Het is goed het onderscheid te maken tussen de steekproefstandaardafwijking en de standaardafwijking gebaseerd op gegevens uit een volledige onderzoekspopulatie. Details horen eigenlijk niet in dit hoofdstuk thuis, maar we zullen er in de les wel op terugkomen.

### **De gemiddelde absolute afwijking**

Een klassieker, die echter niet veel meer gebruikt wordt, is de **gemiddelde absolute afwijking**.

Dit is de som van de *absolute waarden* van de afwijkingen van elke waarde ten aanzien van het rekenkundig gemiddelde, gedeeld door het aantal waarnemingen. Men gebruikt de absolute waarden, omdat de som van alle (positieve en negatieve) afwijkingen tegenover een gemiddelde altijd nul wordt.

### **De variatie**

De **variatie** of nog de “**Sum of Squares**” (**afgekort: SS of soms ook var**) genoemd, is de som van de gekwadrateerde afwijking van elke waarde tegenover het gemiddelde.

$$SS = \sum (X - \bar{X})^2 \quad \text{of} \quad SS = \sum_{i=1}^N (x_i - \bar{x})^2$$

Men **kwadrateert** de verschillen tussen elke geobserveerde waarde van een metrische variabele X en het gemiddelde omdat de som van de afwijkingen van elk individu ten opzichte van het gemiddelde altijd nul is. Kwadrateren lost dit probleem op. De optelsom van de kwadraatafwijkingen geeft een indicatie van spreiding: hoe groter de maat, hoe groter de verschillen tussen de statistische eenheden. Bij een constante zal er nooit spreiding zijn, en dus zullen alle spreidingsmatten nul zijn. De waarde nul betekent dus: iedereen heeft een gelijke score.

### **De (steekproef)variantie**

De **variantie** ( $s^2$ ) is de variatie gedeeld door het aantal onderzoekseenheden wanneer we over populatiegegevens beschikken en gedeeld door het aantal steekprofeenheden minus één.<sup>5</sup> In dit handboek wordt steeds gewerkt met de formule van de *steekproefvariantie*. De formule voor de berekening van de variantie uit een steekproef ziet er als volgt uit:

---

<sup>5</sup> Dit noemen we het aantal vrijheidsgraden (degrees of freedom) wanneer we ons baseren op gegevens verkregen uit een steekproef. Het aantal vrijheidsgraden van een statistiek kan algemeen gesproken het aantal ongebonden variabelen worden genoemd. Of nog: het aantal variabelen die verschillende waarden kunnen aannemen of variëren. Voor het berekenen van de spreiding van een populatie is het aantal vrijheidsgraden gelijk aan het aantal van deze populatie. In het geval van een steekproef verliezen we één vrijheidsgraad. Statistische verwerkingspakketten berekenen trouwens ook de variantie en standaardafwijkingen van kenmerken op basis van de formule die we toepassen op steekproeven.

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

### **De (steekproef)standaardafwijking**

De **steekproefstandaardafwijking** is de vierkantswortel van de **steekproefvariantie**. De **populatiestandaardafwijking** is de vierkantswortel van de **populatievariantie**. In deze syllabus hanteren we steeds de formule voor de steekproefstandaardafwijking en gebruiken we hiervoor soms de afkorting “**std**” van het Engelse begrip “**standard deviation**”.<sup>6</sup>

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \text{ of } s = \sqrt{s^2}$$

Uit de wiskundige notatie kunnen we één en ander afleiden: bereken voor elke meetwaarde de afstand tot het gemiddelde. Neem het kwadraat hiervan voor elke waarde. Tel deze bij elkaar op en deel ze door het aantal observaties (of onderzoekseenheden) min één. De vierkantswortel (square root) is nodig om ook de negatieve afstanden correct mee te tellen.

## **5. Zelf uitrekenen van gemiddelde, variantie en standaardafwijking**

Hoewel het zelf uitrekenen van statistische coëfficiënten in het beroepsleven gebeurt aan de hand van statistische verwerkingspakketten, is het toch cruciaal dat het gemiddelde, de variantie en standaardafwijking nog steeds kunnen berekend worden. Deze vormen immers de belangrijkste parameters die de basis vormen voor vele meer geavanceerde vormen van statistische analyses. Als u de berekening zelf uitvoert, ziet u wat er gebeurt achter de schermen van een statistische analyse. U begrijpt de uitkomst beter omdat u weet hoe u aan de uitkomst bent geraakt. We tonen hieronder een tabel die informatie verschafft over vijftien studenten. Van elke student is hun score op T1 (toets 1) genoteerd. Deze score is niet voor iedereen gelijk. De test staat op 50 punten. An heeft dus 30 op 50, Arno heeft 45 op 50, enz.. We stellen vast dat sommige studenten het heel goed hebben gedaan, maar er zijn toch studenten die het wat minder hebben gedaan. Het rekenkundig gemiddelde is 30. Hoe varieert nu elke student tegenover dat

---

<sup>6</sup> We wijzen hier op het belang van een perfect begrip van het statistische concept ‘standaardafwijking’. Een goed begrip ervan is de basis voor het standaardiseren van waarnemingen (i.e. het omzetten van scores naar z-scores) in het volgende hoofdstuk 4 (de standaardnormale verdeling en diens eigenschappen). Een z-score is immers niets anders dan het *aantal* standaardafwijkingen dat een bepaalde score boven of onder het gemiddelde van de reeks scores ligt.

gemiddelde? Laten we eens kijken. An wijkt niet af van het gemiddelde. Haar deviatiescore is nul. Arno wijkt wel af van het gemiddelde. Hij presteert beter. Zijn deviatiescore is 15. Hiermee zit hij fel boven het gemiddelde. Als we nu voor iedereen de deviatiescore hebben berekend, dan kunnen we de som nemen van die deviatiescores. Die is nul. Dat mag geen verassing zijn, want het is precies een eigenschap van het rekenkundig gemiddelde, met name dat de som van diens afwijkingen steeds nul bedraagt.

Precies omdat de som van alle afwijkingen tegenover het rekenkundig gemiddelde nul bedraagt, kunnen we de deviatiescores niet gebruiken om de spreiding of afwijking rond het rekenkundig gemiddelde uit te drukken. Daar is wat op gevonden. In de kolom ernaast ziet u dat we het kwadraat hebben berekend van elke afwijking tegenover het rekenkundig gemiddelde. Deze oplossing is bijzonder eenvoudig en toch geniaal: de waarden verschillen van nul en men kan de som nemen van al deze waarnemingen. Die som is nu niet meer gelijk aan nul.

Hoewel dit zeer handig is, voelt u wellicht intuïtief aan dat er iets vervelends gebeurt: grote afwijkingen worden uitvergroot door te kwadrateren. Dat is juist. Daarom heeft men hier ook iets op gevonden: we nemen gewoon die som en delen deze door het aantal waarnemingen, of nog beter, we nemen van deze laatste maat gewoon de vierkantswortel. Dit zorgt voor de correctie op de uitvergroting van verschillen. Reken even mee:

**Tabel: studentscores, deviatiescores en het kwadraat daarvan**

Student	ScoreT1	$x_i - \bar{x}$	$(x_i - \bar{x})^*$ $(x_i - \bar{x})$
An	30,00	0	0
Arno	45,00	15	225
Bart	35,00	5	25
Björn	20,00	-10	100
Delphine	40,00	10	100
Hanne	35,00	5	25
Henk	30,00	0	0
Ines	30,00	0	0
Jeroen	25,00	-5	25
Jurgen	20,00	-10	100
Kim	40,00	10	100
Robert	25,00	-5	25
Nele	20,00	-10	100
Sara	25,00	-5	25
Sofie	30,00	0	0
N= 15 $\bar{x} = 30$			Sum of squares 850

$Var X = \text{Sum of squares } X/(N-1) = 850/14 = 60.71$

$Std X = \sqrt{60.71} = 7.79$

### Werkwijze :

- Stap 1:** Bereken het rekenkundig gemiddelde van de variabele
- Stap 2:** Trek het rekenkundig gemiddelde af van iedere waarde (kolom 2)
- Stap 3:** Kwadrateer de verschillen: dit gebeurt omdat anders positieve verschillen tegen negatieve zouden wegvalLEN. (het totaal van de afwijkingen zou dan nul zijn!)
- Stap 4:** Tel de gekwadrateerde verschillen op.
- Stap 5:** Bereken de variantie: deel het totaal van de gekwadrateerde verschillen door het aantal waarnemingen min één.
- Stap 6:** Bereken de standaardafwijking: trek de wortel uit de variantie. De standaardafwijking wordt hierdoor weer in vergelijkbare hoeveelheden gegeven als het gemiddelde.

### *De variatiecoëfficiënt*

Het nadeel van hiervoor besproken spreidingsmatten is hun afhankelijkheid van de meeteenheid. Als we nu twee totaal verschillende variabelen beschouwen: criminaliteit en IQ. Beide worden gemeten in een andere meeteenheid. Je kan van elke variabele de standaardafwijking berekenen. Je zal ongetwijfeld vaststellen dat het ene kenmerk een grote standaardafwijking heeft dan het andere kenmerk. Dat is juist, maar dat kan heel misleidend zijn. Want: precies omdat de beide kenmerken in een eigen meeteenheid zijn uitgedrukt, kunnen we die toch niet gaan vergelijken met elkaar? We riskeren een kanjer van een fout te begaan als we dat zouden doen. Voor dat probleem hebben statistici een handige oplossing bedacht. Aangezien beide kenmerken verschillen in termen van meeteenheid, en dus een eigen gemiddelde hebben, delen we gewoon de standaardafwijkingen door de respectieve gemiddeldes. Het resultaat is de variatiecoëfficiënt v. De **variatiecoëfficiënt (v)** is een **gestandaardiseerde spreidingsmaat**. Wanneer we de spreiding van de prijs van een aantal producten willen weten en hiervoor de standaardafwijking berekenen van de prijs uitgedrukt in Belgische frank en daarna uitgedrukt in Euro, dan zal de bekomen maat verschillend zijn, terwijl ze in feite op dezelfde gegevens berust. Om die reden werd de variatiecoëfficiënt ingevoerd. Men zegt wel eens van de variatiecoëfficiënt dat deze **dimensieloos** (niet afhankelijk van de meeteenheid) is en dit laat toe de spreiding van

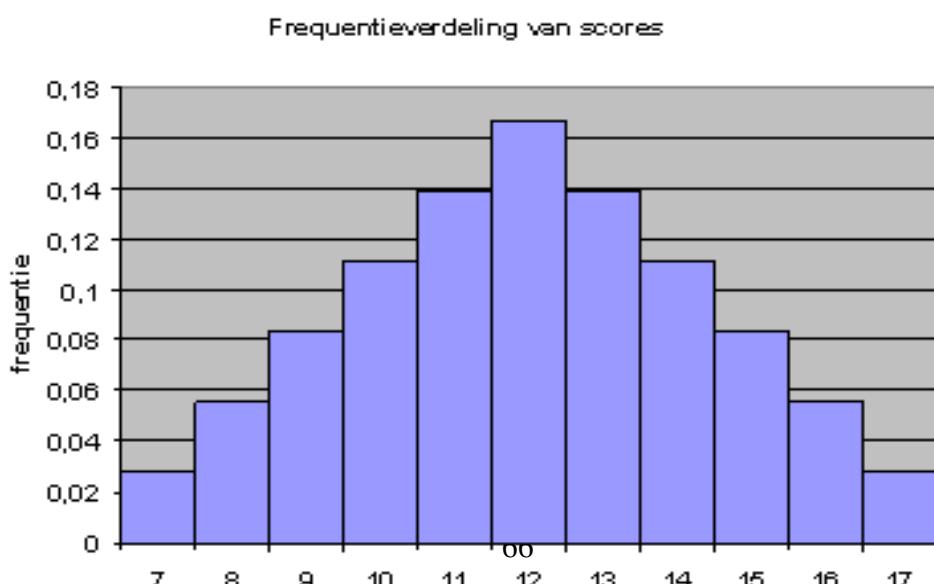
verdelingen die worden uitgedrukt in *verschillende meeteenheden te vergelijken*. De variatiecoëfficiënt wordt berekend door de standaardafwijking te delen door het rekenkundige gemiddelde (beide worden immers in dezelfde meeteenheid uitgedrukt). Onthoud deze belangrijke eigenschap van de variatiecoëfficiënt.

$$v = \frac{s}{\bar{X}}$$

## 6. Parameters van vorm

Naast centraliteit en spreiding kunnen we ook de **vorm** van de *verdeling van kenmerken* samenvatten aan de hand van enkele parameters. Vergelijken we delingen inzake vorm, dan kunnen we vaststellen dat delingen verschillen in de mate waarin zij afwijken van een *symmetrische verdeling*. Bij het bestuderen van de symmetrie van een verdeling bekijken we in feite hoe de gegevens verdeeld zijn ten opzichte van het rekenkundig gemiddelde. Uit dit laatste kunnen we reeds afleiden dat symmetrie enkel kan bestudeerd worden voor **metrische meetsschalen**. Een verdeling kan symmetrisch, links asymmetrisch of rechts asymmetrisch zijn. Laten we dit illustreren aan de hand van enkele voorbeelden. Een verdeling is **symmetrisch** als het rekenkundig gemiddelde en de mediaan aan elkaar gelijk zijn. Een symmetrisch verdeelde variabele is de normaal verdeelde variabele. Dit is een variabele die de gekende Gauss-curve volgt. De Gauss-curve werd door de statisticus Quetelet ontdekt voor sociale kenmerken.

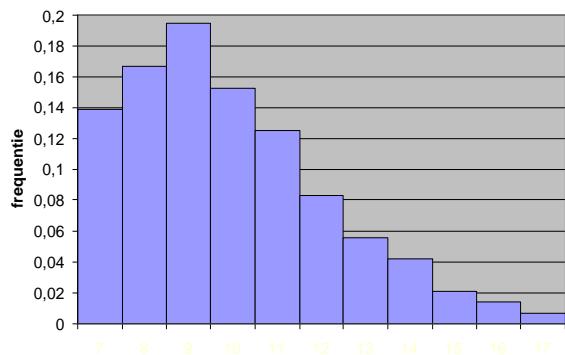
Een symmetrische verdeling ziet er als volgt uit:



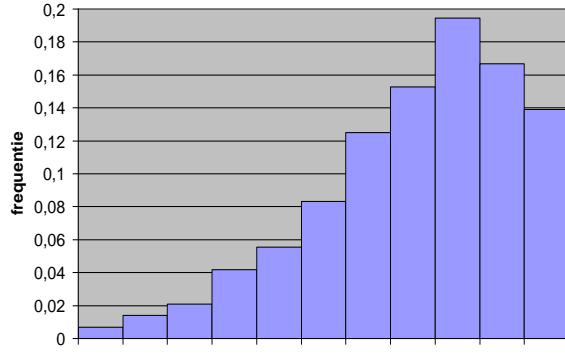
Deze symmetrische verdeling noemen we symmetrisch omdat van het feit dat de linkerhelft en de rechterhelft perfect op elkaar lijken. Ze zijn elkaar spiegelbeeld.

Een verdeling is **positief asymmetrisch** als de verdeling een langere staart naar rechts heeft. In dit geval is het rekenkundig gemiddelde groter dan de mediaan. Positieve asymmetrie betekent dat de hogere waarden minder vaak voorkomen dan de lagere waarden. Dit is heel vaak het geval bij criminaliteitsmaten. Vele jongeren hebben geen enkel delict gepleegd in een periode van 12 maanden, een niet onaanzienlijk deel van de jeugd pleegt wel eens een delict, maar slechts heel weinig jongeren plegen heel veel delicten. Een verdeling is **negatief asymmetrisch** als de verdeling links een langere staart heeft. In dat laatste geval komen de lage waarden minder voor dan de hoge waarden.

### Positieve asymmetrie



### Negatieve asymmetrie



De vraag is nu: kunnen we de symmetrie die een variabele kenmerkt ook uitdrukken op basis van een eenvoudig getal, zodat we in een oogopslag, ook zonder naar de figuur te kijken kunnen aflezen hoe symmetrisch een variabele is. Het antwoord hierop is ja. We spreken dan van **parameters van vorm**. Er bestaan verschillende statistische parameters van vorm. Dit zijn de parameters die we nodig hebben om de symmetrie te berekenen. We beperken ons hieronder tot de **empirische coëfficiënt van Pearson**.

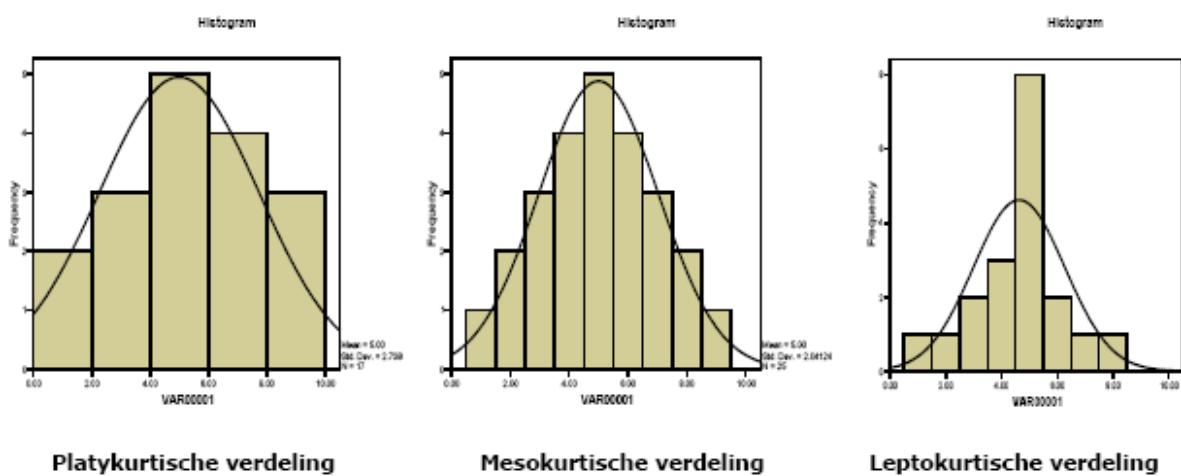
De **empirische coëfficiënt van Pearson** wordt als volgt berekend: (1) bereken het verschil tussen het gemiddelde en de mediaan en (2) deel deze waarde door de standaardafwijking.

$$S = \frac{\bar{X} - \tilde{x}}{s}$$

**Een verdeling is positief asymmetrisch als de coëfficiënt een positieve waarde heeft en negatief asymmetrisch als de coëfficiënt een negatieve waarde heeft.**

Naast de symmetrie kan men ook de mate van **afplatting** van een verdeling bestuderen. De afplatting of **kurtosis** is de mate van afplatting van de gegevens rondom het rekenkundige gemiddelde. De kurtosis wordt steeds met de **standaardnormale** of **Gauss-verdeling** als standaard vergeleken. We besteden verderop in deze syllabus meer aandacht aan deze verdeling. We onderscheiden, **mesokurtische** verdelingen (gemiddelde afplatting), **leptokurtische** verdelingen (scherper) en **platykurtische** verdelingen (platter). Een platykurtische verdeling is een verdeling die lijkt op een normaal verdeeld kenmerk waar iemand spreekwoordelijk met een hamer op heeft geslagen. Daardoor ziet deze er platter uit. Een leptokurtische verdeling is een verdeling waar de hoogst voorkomende waarde zo veel vaker voorkomt, dat deze er echt uitschiet. Om de vergelijking met de normale verdeling mogelijk te maken, werd deze ook getekend op de histogrammen in kwestie.

### Figuur: de kurtosis



## 7. De box plot

Een overzichtelijke manier om gegevens **vanaf ordinaal meetniveau** visueel voor te stellen, is de **box plot**. Een standaard box plot laat toe om een snelle visuele evaluatie te maken van de informatie die vervat zit in een frequentieverdeling.

**De box plot is een grafiek van de vijf-getallensamenvatting.** De vijf-getallensamenvatting van een verdeling bestaat uit de mediaan M, de kwartieren Q1 en Q3 en de minimale en maximale waarnemingen, genoteerd als:

$$(Q3-Q1)*1.5 \text{ of minimale niet-uitschietende waarde} - Q1 - \text{Mediaan} - Q3 - (Q3-Q1)*1.5 \text{ of maximale niet-uitschietende waarde}^7.$$

**Opgelet: minimum en maximum betekenen hier niet de allerkleinste of allergrootste waargenomen waarde. Ze betekenen respectievelijk de hoogste en laagste NIET-UITSCHIETENDE waarde.**

De combinatie van deze vijf getallen is een snelle manier om een samenvatting te krijgen van zowel het centrum als de spreiding van een variabele. De mediaan beschrijft het centrum van de verdeling, de kwartieren tonen de spreiding van de middelste helft van de gegevens (de centrale 50%), de waarde die overeenstemt met anderhalve keer de interkwartielafstand het minimum en maximum tonen de volledige spreiding van de gegevens. Deze vijf-getallensamenvatting leidt tot een grafische verdeling: de box plot.

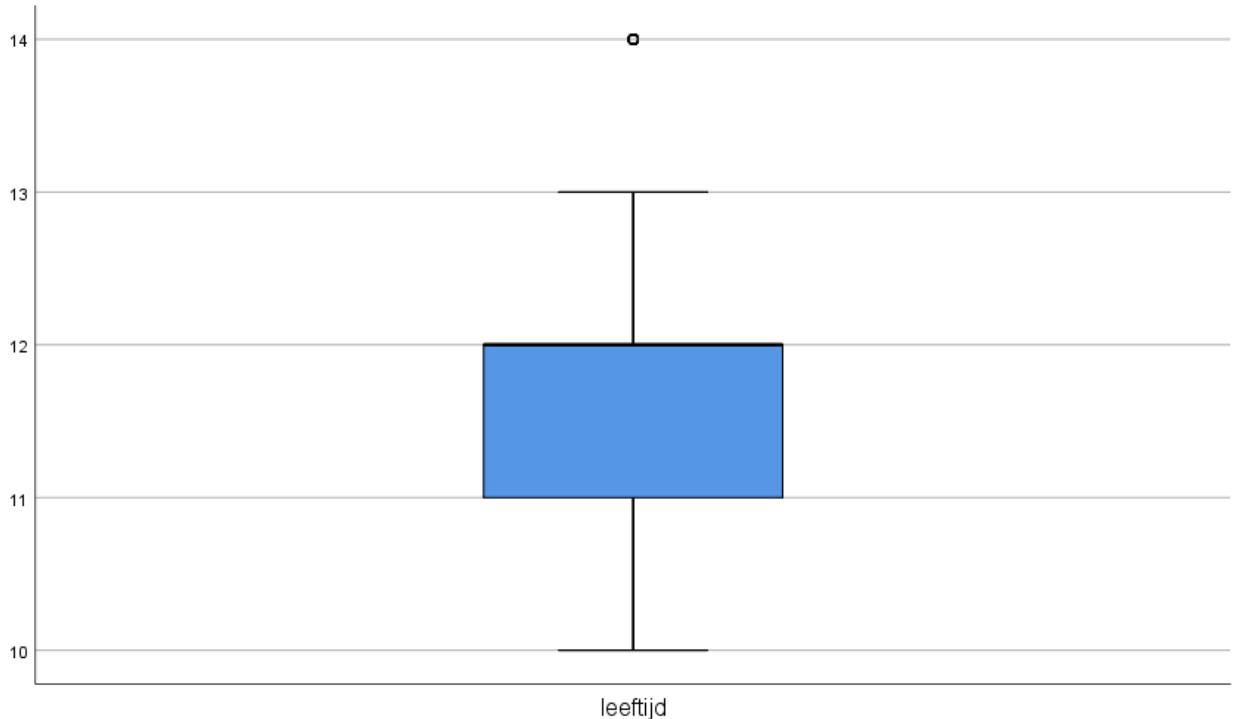
De box plot bestaat uit:

- De centrale box (rechthoek) die zich uitstrek van het eerste kwartiel Q1 tot het derde kwartiel Q3
- De mediaan M die in de rechthoek gemarkeerd wordt door een lijn
- Onder en boven de box zie je twee lijnen (whiskers). Deze tonen de waarden die overeenkomen met anderhalve keer de interkwartielafstand. De waarde die overeenkomt met de onderste en bovenste lijn komt niet noodzakelijk overeen met de allerlaagste en allerhoogste waarneming. Deze komt enkel overeen met de laagste en hoogste “niet-uitschietende” waarneming die zich op de whisker bevindt. Waarnemingen buiten de whiskers zijn dus steeds mogelijk! **Let dus goed op want als je gevraagd wordt de variatiebreedte af te lezen, mag je je niet exclusief door de whiskers laten leiden.** De allerlaagste en allerhoogste waarneming kunnen

<sup>7</sup> Eerder mathematische handboeken gebruiken soms de term minimum en maximum maar bedoelen daarmee de minimale en maximale niet-uitschietende waarde. Om verwarring met het echte minimum in de betekenis van de allerlaagste waarde en echte maximum (maximumwaarde) te vermijden spreken we in dit handboek van minimale niet-uitschietende waarde als waarde die overeenkomt met de onderste whisker en maximale niet-uitschietende waarde als we het hebben over de maximale niet-uitschietende waarde.

immers buiten de whiskers vallen. We spreken dan van outliers of uitbijters. Statistisch gezien komen deze waarden zeer zelden voor, omdat van hun grote afstand tot anderhalve keer de interkwartielafstand (i.e.  $Q3-Q1) * 1.5$ .

*Figuur 1: Box plot van een metrische variabele leeftijd*



De vijf-getallensamenvatting voor de variabele ‘leeftijd’ in *figuur 1* bedraagt:

10 11 12 12 13

Hoe lezen we best een box plot ?

Eerst kijken we naar de mediaan. In *figuur 1* zien we dat de mediaan 12 jaar bedraagt: het is het punt dat precies in het midden valt van de verdeling. Daarna kijk je naar de spreiding. De kwartieren tonen de spreiding van de middelste helft van de gegevens. In bovenstaand voorbeeld is de waarde van  $Q1 = 11$  (=waarde waaronder zich 25% van alle eenheden bevinden) en de waarde van  $Q3 = 12$  (=waarde waaronder zich 75% van alle eenheden bevinden). Bemerk dat in dit voorbeeld de mediaan samenvalt met het derde kwartiel  $Q3$ . We herhalen dat de box zelf de centrale 50% van de eenheden vormt. In dit voorbeeld heeft 50% van de respondenten een leeftijd tussen 11 en 12 jaar.

**De uiterste waarden tenslotte (de kleinste en de grootste waarnemingen) tonen de spreiding van de hele gegevensverzameling. De allerlaagste waarneming valt hier samen met het minimum en bedraagt 10 jaar en het maximum bedraagt 13 jaar. De allerhoogste**

**waarneming is 14 jaar. De variatiebreedte komt overeen met het verschil tussen de allerhoogste en allerlaagste waarneming (14-10= 4).**

Box plots worden in de praktijk gebruikt voor het opsporen van uitzonderlijk lage of hoge scores. Er is hier duidelijk sprake van een uitschieter. Een aantal respondenten hebben een leeftijd van 14jaar.

De variatiebreedte is echter gevoelig voor uitschieters. De afstand tussen de kwartieLEN (=de spreidingsbreedte van de middelste helft van de gegevens) is een meer *resistente* spreidingsmaat<sup>8</sup>. We noemen deze afstand de interkwartielafstand (*IKA*= de afstand tussen het eerste en het derde kwartiel.  $IKA = Q3 - Q1$ ). In ons voorbeeld hierboven geldt dat  $IKA = 12 - 11 = 1$ .

De interkwartielafstand wordt vooral gebruikt als vuistregel voor het opsporen van verdachte uitschieters.

Een waarneming is een verdachte uitschieter als deze ten minste  $1,5 \times IKA$  boven het derde kwartiel of onder het eerste kwartiel ligt. We noemen dit het *1,5 x IKA-criterium* voor uitschieters.

In ons voorbeeld in figuur 1 geldt:

$$1,5 \times IKA = 1,5 \times 1 = 1,5$$

De waarden onder  $11 - 1,5 = 9,5$  en boven  $12 + 1,5 = 13,5$  zijn herkenbaar als mogelijke uitschieters. In ons voorbeeld zijn er geen uitschieters naar beneden maar wel naar boven. Er zijn respondenten met een leeftijd van 14jaar.

Statistische software zoals SPSS maken gebruik van de  $1,5 \times IKA$ . Box plots die door software zijn getekend zijn vaak **gemodificeerde box plots** die verdachte uitschieters afzonderlijk weergeven. In SPSS worden uitzonderlijk lage/hoge scores aangegeven met ° en met een \*.

Waarnemingen aangeduid als ° bevinden zich tussen  $Q1 - 1,5 IKA$  en  $Q1 - 3 IKA$  enerzijds en tussen  $Q3 + 1,5 IKA$  en  $Q3 + 3 IKA$  anderzijds. Dit zijn ‘zwakte’ uitschieters.

Waarnemingen gemarkerd met een \* bevinden zich buiten  $Q1 - 3 IKA$  en  $Q3 + 3 IKA$ .

Deze waarnemingen zijn mogelijks uitschieters die de uitkomsten te sterk beïnvloeden. Het zijn ‘extreme’ uitschieters.

De lijnen die uit de centrale box komen, hebben dan alleen betrekking op kleinste en grootste waarnemingen die niet voldoen aan de  $1,5 \times IKA$ -regel. In ons voorbeeld zijn de waarnemingen

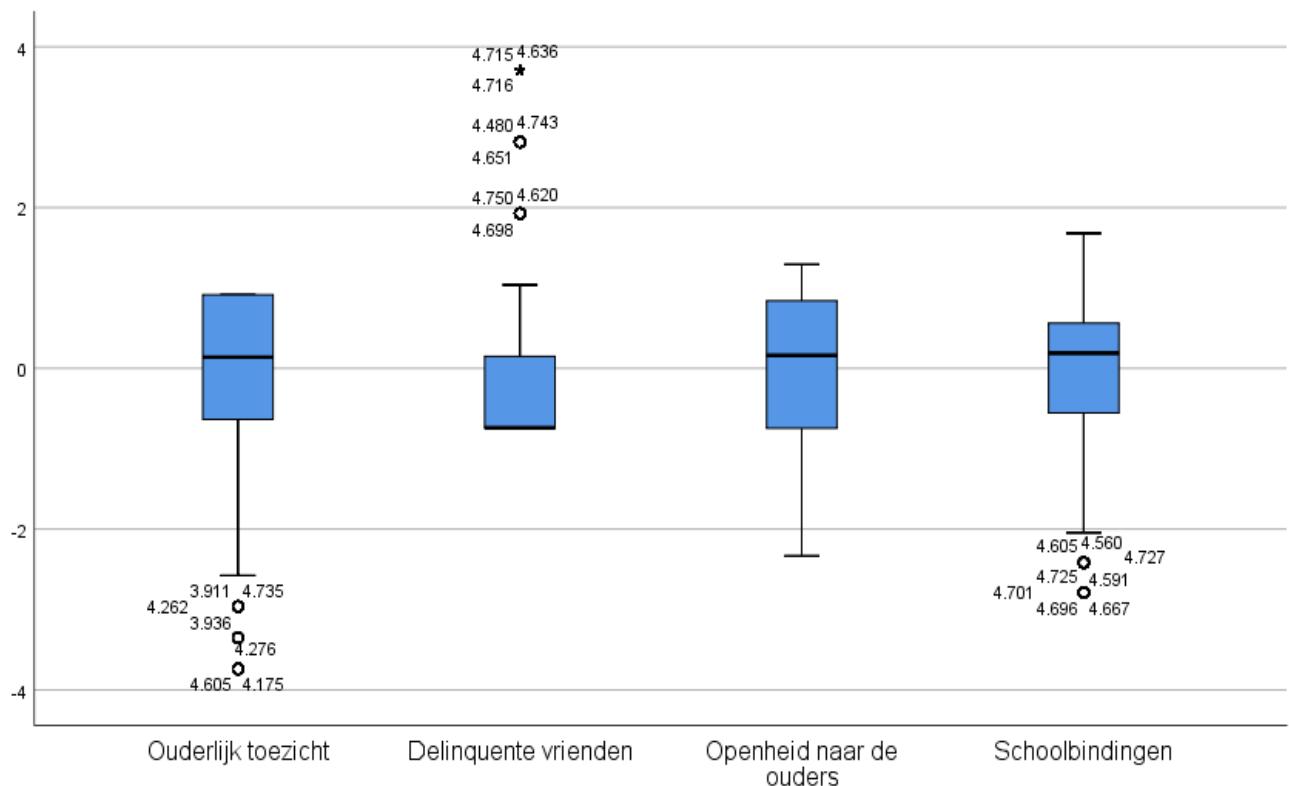
<sup>8</sup> Een resistente maat van een verdeling wordt nauwelijks beïnvloed door veranderingen in de numerieke waarden van een klein gedeelte van het totale aantal waarnemingen. Bijvoorbeeld: de mediaan en kwartieLEN zijn resistente maten, maar het gemiddelde en de standaardafwijking zijn dat niet. We komen hier verder in de cursus op terug.

die hieraan niet voldoen respectievelijk 10 en 13 en zijn er uitschieters naar boven met een waarde 14.

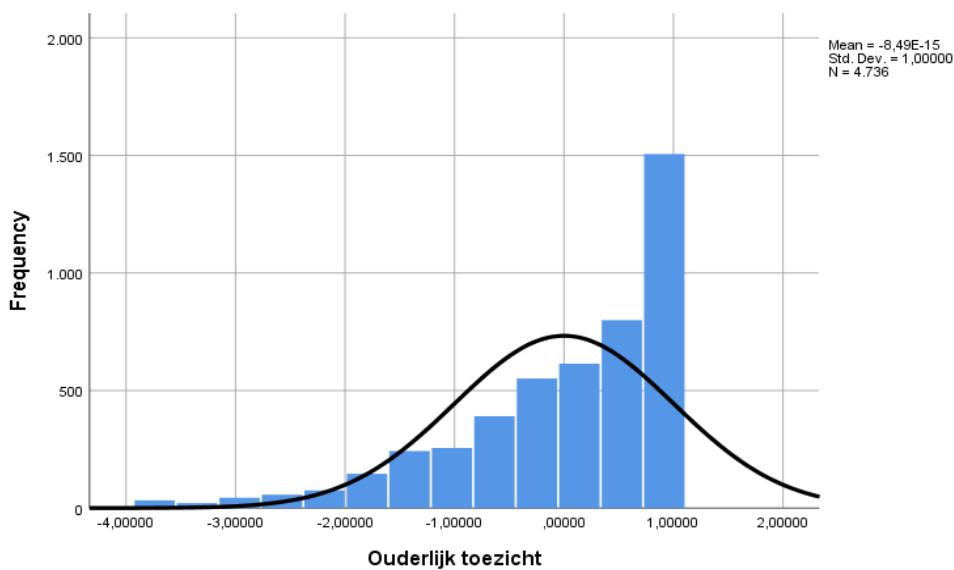
Box plots detecteren niet enkel uitschieters. Omdat box plots minder details laten zien dan bijvoorbeeld een histogram, worden ze bij voorkeur gebruikt om verschillende verdelingen met elkaar te vergelijken zoals afgebeeld in *figuur 2*. Figuur 2 toont de afzonderlijke box plots voor 4 variabelen ‘ouderlijk toezicht’, ‘delinquente vrienden’, ‘openheid naar de ouders’ en ‘schoolbindingen’. De variabelen werden gemeten door aan respondenten verschillende vragen te stellen en hun scores op elke vraag samen te tellen. De variabelen werden gestandaardiseerd om de interpretatie eenduidig te maken.

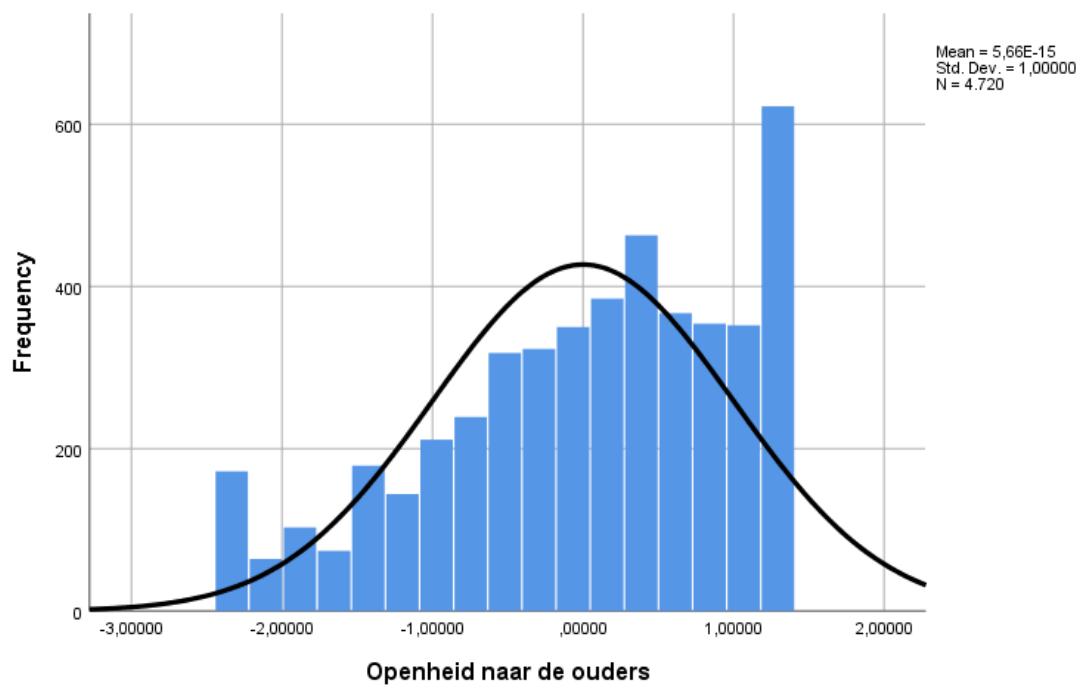
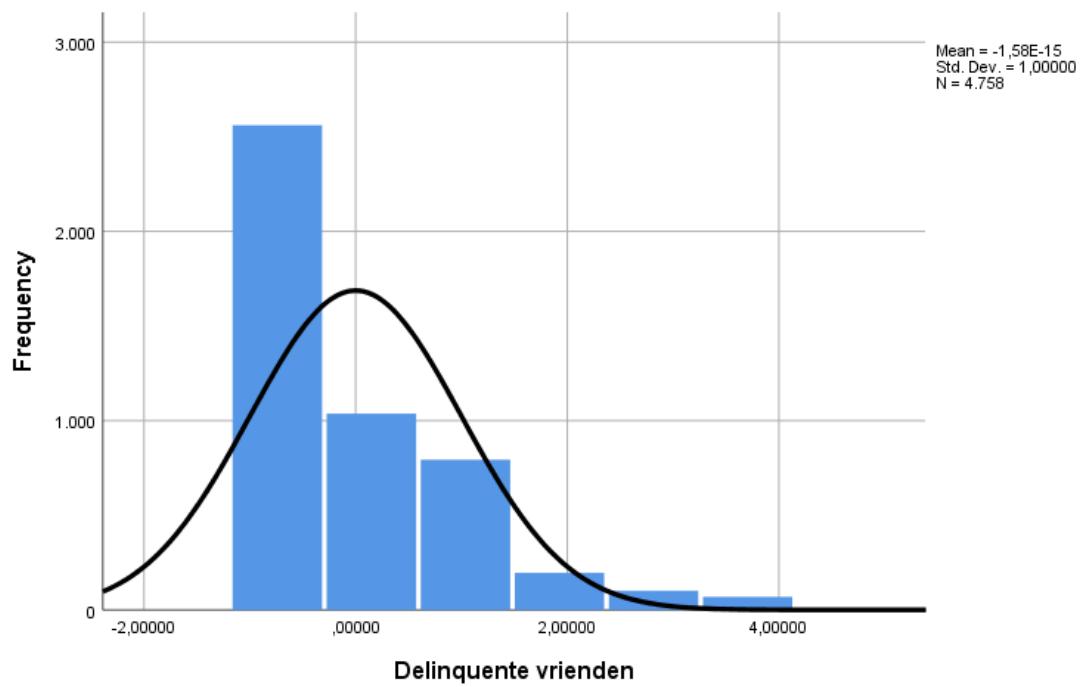
Kijken we eerst naar de mediaan, dan zien we dat deze het laagst is voor de variabele ‘Delinquente vrienden’. Kijken we naar de kwartieLEN die ons de spreiding tonen van de middelste helft van de verdeling, dan zien we dat deze het kleinst is voor de variabele ‘Delinquente vrienden’. Kijken we naar de uiterste waarden of de kleinste en de grootste waarnemingen dan zien we voor de variabele ‘openheid naar de ouders’ geen uitschieters. Er zijn geen respondenten die op dit kenmerk hoger of lager scoorden dan  $1,5 \times IKA$  en  $3 \times IKA$ . Er zijn wel uitschieters voor de variabelen ‘ouderlijk toezicht’ en ‘schoolbindingen’, uitschieters naar beneden, en voor de variabele ‘delinquente vrienden’ zijn er uitschieters naar boven.

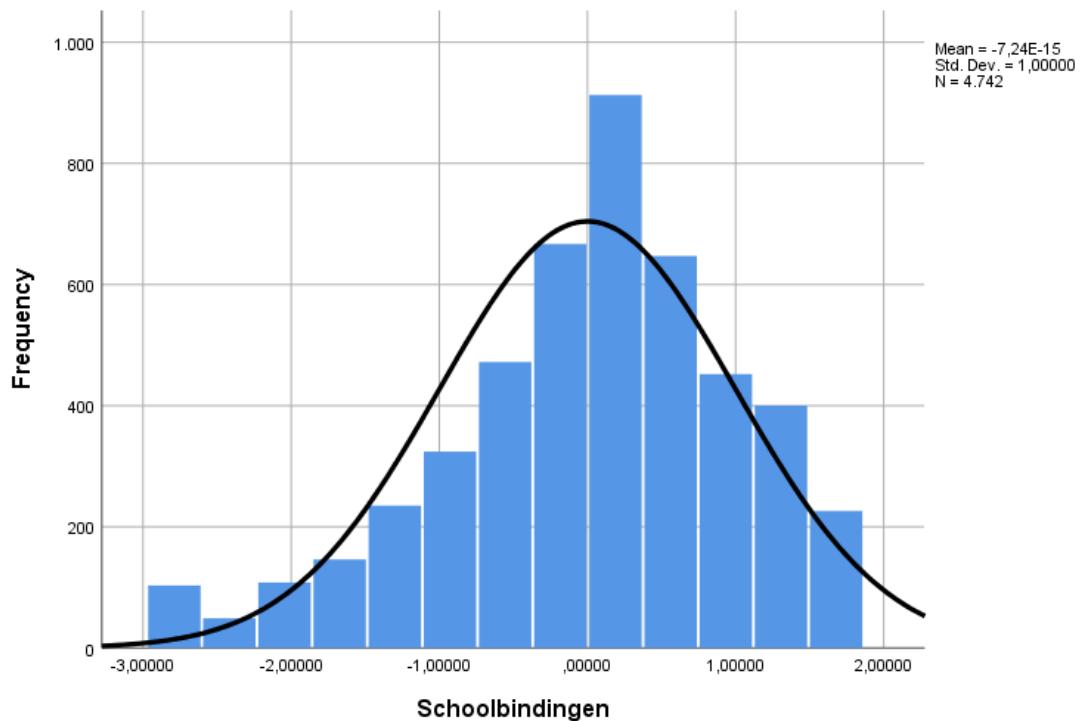
*Figuur 2: Box plots voor ouderlijk toezicht, delinquenten vrienden, openheid naar de ouders en schoolbindingen.*



Deze box plots zeggen ons ook iets over de vorm van de verdelingen. De variabelen ‘ouderlijk toezicht’ en ‘schoolbindingen’ hebben nogal wat uitschieters naar beneden en zijn links scheef verdeeld. De variabele ‘openheid naar de ouders’ is links scheef verdeeld. De variabele ‘delinquenten vrienden’ is rechts scheef verdeeld. Ter controle en vergelijking presenteren we hierna tenslotte het histogram met aanduiding van de normaalcurve van elke variabele.







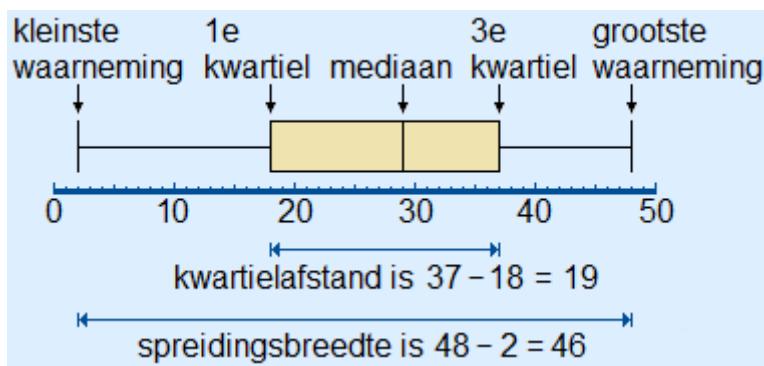
## Samenvatting

De **vijf-getallensamenvatting**, bestaande uit de mediaan, de kwartieren en minimale en maximale niet-uitschietende waarde, verschaft een snelle en globale beschrijving van een verdeling. De mediaan beschrijft het centrum en de kwartieren en de whiskers worden gebruikt om aan te geven wat statistisch gezien abnormaal is: waarden die daarbuiten vallen, komen uiterst zelden voor.

De **interkwartielafstand (IKA)** is het verschil tussen de kwartieren. Het is de spreiding van de middelste helft van de gegevens. Het  $1,5 \times IKA$ -criterium merkt waarnemingen die ten minste  $1,5 \times IKA$  voorbij de kwartieren vallen als mogelijke uitschieters.

**Box plots** die berusten op de vijf-getallensamenvatting zijn nuttig om verschillende verdelingen met elkaar te vergelijken. De centrale rechthoek strekt zich uit van het eerste tot het derde kwartiel en geeft de spreiding weer van de middelste helft van de verdeling. De mediaan wordt in de centrale rechthoek gemarkeerd. Lijnen strekken zich uit tot aan de kleinste en de grootste waarneming en geven de volledige spreiding van de gegevens weer, behalve de punten gemerkt door het  $1,5 \times IKA$ - criterium, deze worden vaak afzonderlijk weergegeven als uitschieters.

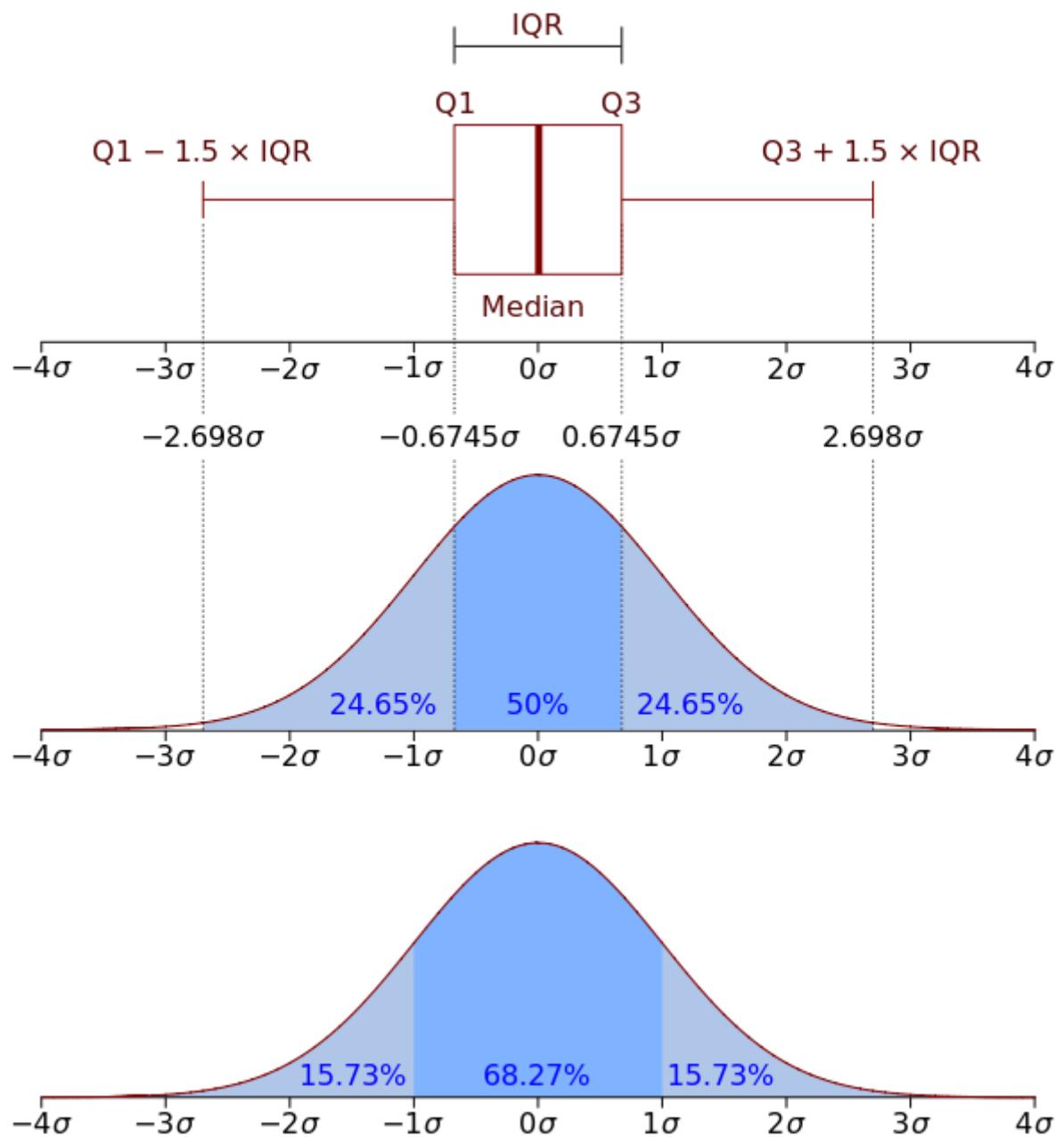
## Samenvattende figuur



We herhalen dat in deze cursus onder ‘kleinste waarneming’ wordt verstaan : minimale niet-uitschietende waarde of  $(Q3 - Q1) * 1,5$  en onder ‘grootste waarneming’: maximale niet-uitschietende waarde of  $(Q3 - Q1) * 1,5$ .

De vijf-getallensamenvatting is niet de meest gangbare numerieke beschrijving van een verdeling. Een geschikter middel hiervoor is een combinatie van het gemiddelde om het centrum te meten en de standaardafwijking om de spreiding te meten. Ter herhaling: de standaardafwijking meet de spreiding door te kijken hoe ver de waarnemingen van hun gemiddelde zijn verwijderd. De standaardafwijking  $s$  is de wortel van de variantie  $s^2$ . Het gemiddelde en de standaardafwijking zijn goede beschrijvingen voor symmetrische delingen zonder uitschieters. Ze zijn bijzonder nuttig voor de normale en standaardnormale delingen die in het volgende hoofdstuk aan bod komen. In *figuur 3* presenteren we een vergelijking van een box plot met een standaardnormale verdeling en leggen hiermee de link als overgang naar het volgende hoofdstuk.

Figuur 3: vergelijking van een box plot met een standaardnormale verdeling



## **8. Testvragen**

Hieronder vind je enkele uitspraken over de univariate statistiek. Deze vragen kan je gebruiken om je parate kennis te toetsen over de basiskennis die tot nog toe werd meegegeven. **In gewijzigde vorm kunnen dergelijke theorievragen ook op het examen voorkomen.** Deze vragen zijn afkomstig uit vroegere examens. De correcte antwoorden zijn rechtstreeks uit de cursus afleidbaar. De antwoorden vind je achteraan dit theorieboek. Dit is niet meer dan een test om te zien of je mee bent met de leerstof. Faal je op deze testvragen, dan is het hoogdringend tijd om in actie te schieten.

### **1. Deviatiescores zijn**

- De som van de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde
- De som van de afwijkingen tegenover het rekenkundig gemiddelde

### **2. De mediaan is een robuuste parameter van centraliteit**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

### **3. De Index van Diversiteit kan geïnterpreteerd worden als de kans dat twee willekeurig gekozen onderzoekseenheden tot een verschillende categorie behoren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

### **4. De variatieratio neemt de waarde van nul aan indien alle waarnemingen tot de modale categorie behoren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

### **5. Een variabele van het metrisch niveau kan bestudeerd worden op ordinaal niveau**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**6. Een variabele van het ordinale niveau kan bestudeerd worden op het metrische niveau**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**7. De variatiebreedte is het verschil tussen de maximale waarde en de minimale waarde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**8. De variatiebreedte is de som van de maximale en minimale waarde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**9. Uitspraken extrapoleren van de steekproef naar de populatie doe je via**

- de beschrijvende univariate statistiek
- de inferentiële statistiek

**10. De steekproefstandaardafwijking wordt berekend op basis van de formule van de populatiestandaardafwijking maar in de noemer staat  $n+1$**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**11. Een bar chart wordt gebruikt**

- Voor variabelen op het metrische niveau
- Voor variabelen op het categorische niveau

**12. Variabelen die op een histogram worden gepresenteerd zijn steeds ratio niveau**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**13. Een lijndiagram wordt gebruikt voor de**

- Visuele voorstelling van metrische gegevens die niet in klassen zijn gegroepeerd (ruwe scores)
- Visuele voorstelling van metrische gegevens die wel in klassen zijn gegroepeerd

**14. Een frequentiepolygoon wordt gebruikt voor de**

- Visuele voorstelling van metrische gegevens die niet in klassen zijn gegroepeerd (ruwe scores)
- Visuele voorstelling van metrische gegevens die wel in klassen zijn gegroepeerd

**15. Een platykurtische variabele is**

- Platter dan een normaal verdeelde variabele
- Scherper dan een normaal verdeelde variabele

**16. Een box-plot kan worden gebruikt voor variabelen vanaf**

- Het nominaal niveau
- Ordinaal niveau
- Interval niveau
- Ratio niveau

**17. De mediaan komt overeen met**

- Het vijftigste percentiel
- Het eenenvijftigste percentiel

**18. Als we een kenmerk dat perfect normaal verdeeld is voorstellen via een box-plot, dan is de afstand tussen de mediaan en de hoogste waarde even groot als de afstand tussen de mediaan en de laagste waarde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**19. Een variabele die rechtsscheef verdeeld is**

- Heeft een langere staart naar rechts
- Heeft een langere staart naar links
- Heeft geen staart

**20. De interkwartielafstand is een maat van**

- Centraliteit
- Spreiding
- Vorm

**21. Een frequentieverdeling kunnen we opvatten als een kansverdeling**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**22. Een onderzoekseenheid heeft een z-score van -2.20 voor het metrisch kenmerk “studieresultaat” .**

- De onderzoekseenheid valt buiten de centrale 95% van de waarnemingen.
- De onderzoekseenheid doet het beduidend beter dan de gemiddelde onderzoekseenheid

**23. Centreren wil zeggen dat men een kenmerk uitdrukt als een afwijking tegenover het rekenkundig gemiddelde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**24. De frequentieverdeling (histogram) verandert vormelijk niet wanneer men standaardiseert**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**25. Het rekenkundig gemiddelde is een spreidingsmaat die gevoelig is voor uitschieters**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**26. Variabelen van het categorische niveau bevatten categorieën. Een onderzoeksseenheid mag tegelijkertijd in twee categorieën van dezelfde variabele zitten**

- Dit mag niet als we de regels van de statistiek volgen
- Dit mag wel, er zijn hier geen regels voor

**27. Operationalisering betekent**

- Dat we een kenmerk meetbaar maken
- Dat we een kenmerk van een conceptuele definitie voorzien

**28. De populatievariantie is de som van de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde, gedeeld door het steekproefeffectief**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**29. De variatieratio is een spreidingsmaat die we moeten gebruiken wanneer twee metrische kenmerken in een verschillende eenheid werden gemeten en men wil de spreiding vergelijken**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**30. De keuze voor een univariate parameter wordt ingegeven door**

- De onderzoeksvergadering
- Het meetniveau
- Beide

**31. Stel: je bent activist die ijvert voor meer inkomensgelijkheid. Om de inkomensongelijkheid te demonstreren, kan je kiezen tussen een aantal parameters. Inkomen is zeer scheef verdeeld.**

- Je presenteert vanuit je standpunt de mediaan
- Je presenteert vanuit je standpunt het rekenkundig gemiddelde

## **9. Leerdoelen**

Het hoofdstuk over de univariate beschrijvende statistiek is een belangrijk hoofdstuk. De kennisopbouw in dit hoofdstuk is cumulatief. Je zal gemerkt hebben dat je de elementaire kennis en basisbegrippen uit het vorige hoofdstuk dient te begrijpen vooraleer je de essentie van de univariate statistiek kent en kan toepassen. Univariate statistiek wordt gebruikt om onderzoeks vragen te beantwoorden. Het gaat om univariate beschrijvende onderzoeks vragen die peilen naar de centrale tendensen, naar spreiding en naar vorm.

Het is belangrijk dat je volgende begrippen eigen maakt: absolute frequenties, proporties, relatieve frequenties en cumulatieve absolute en relatieve frequenties. Het is belangrijk dat je inziet dat het meetniveau van een variabele heel belangrijk is met betrekking tot de keuze voor een beschrijvende analysetechniek. Als onderzoeker heb je de vrijheid. Gebruik die en baseer je daarbij op wat je weet. Het indelen in klassen is theoretisch belangrijk, maar wordt in de criminologische praktijk vaak gedaan aan de hand van informatieverwerkingspakketten en op basis van kwartieren en meer inhoudelijke gronden.

We hebben een aantal manieren voorgesteld om statistische gegevens grafisch voor te stellen. De belangrijkste zijn het taartdiagram en het staafdiagram voor nominale en ordinale kenmerken. Bedenk dat het aantal categorieën best beperkt is bij taartdiagrammen. In een staafdiagram kunnen meer categorieën worden voorgesteld op een duidelijke manier dan in een taartdiagram. Voor de cumulatieve voorstelling van gegevens kunnen we het cumulatief frequentiediagram gebruiken. Metrische gegevens worden aan de hand van een histogram voorgesteld. Andere manieren om metrische gegevens voor te stellen zijn lijndiagrammen en frequentiepolygonen. We komen ze echter minder vaak tegen dan het histogram. Wees steeds alert bij grafieken. De manier waarop gegevens in assen worden voorgesteld kan al even

misleidend als verhelderend zijn. Uitspraken zoals “met statistiek kan je alles bewijzen” zijn partieel gebaseerd op misleidende voorstellingen.

Het meetniveau van een variabele is ook bepalend voor de keuze van een beschrijvende analysetechniek. Je moet heel goed het onderscheid kennen tussen de parameters van centraliteit, spreiding en vorm. Je moet heel goed beseffen welke parameters kunnen gehanteerd worden en daarbij moet je steeds de link leggen met het meetniveau van het bestudeerde kenmerk.

De parameters van centraliteit die belangrijk zijn in criminologisch onderzoek zijn de modus, de mediaan, de kwantielen en het rekenkundig gemiddelde. Elk geven zij op hun manier de centrale tendensen weer. Je moet weten hoe je deze dient te berekenen en je moet deze kunnen interpreteren. Je moet de eigenschappen van het rekenkundig gemiddelde zeer goed kennen.

De parameters van spreiding die we gezien hebben bij de studie van spreiding op het nominale niveau zijn de variatieratio, de index van diversiteit. Deze moet je zelf kunnen berekenen. Vanaf het ordinale niveau hebben we gezien dat de variatiebreedte en de interkwartielafstand interessante spreidingsmaten zijn. Deze moet je uit tabellen kunnen afleiden. De parameters van spreiding die gehanteerd worden op het metrische niveau zijn zeer belangrijk en zullen steevast terugkeren in deze syllabus. Deze moet je dus zeer goed begrijpen. Je moet het onderscheid kennen tussen de gemiddelde absolute afwijking, de variatie, de variantie, de standaardafwijking en de variatiecoëfficiënt. Je dient deze zelf te kunnen berekenen en je dient de uitkomsten van univariate analyses zelf inhoudelijk te kunnen interpreteren.

Tot slot hebben we de parameters van vorm gezien. We hebben aandacht besteed aan symmetrie en kurtosis. Je dient te weten wat symmetrie betekent en hoe afwijkingen ten opzichte van symmetrie kunnen worden vastgesteld. Je dient natuurlijk wel te weten wat kurtosis is, maar we hechten in deze starterscursus meer belang aan asymmetrie. We hebben ook gezien dat we een kenmerk kunnen voorstellen aan de hand van een box plot. Je dient te weten wat je kan afleiden uit een box plot.

In concreto betekent dit dat je niet enkel kennis dient te hebben over de centrale begrippen uit de univariate beschrijvende statistiek, maar dat je ook inzichten dient te hebben gekomen. Je dient correct enkele tabellen en beschrijvende coëfficiënten te kunnen interpreteren en steeds

de link te kunnen leggen met de onderzoeksvergadering die achter het gebruik van een beschrijvende statistische parameter gaat.

Hoe bereid je dit hoofdstuk best voor met betrekking tot het examen? Het examen bestaat uit meerkeuzevragen. Voor concrete voorbeelden verwijzen we naar deel II van dit handboek, met name de praktische oefeningen en voorbeelden. Je zal situaties voorgelegd krijgen, zoals tabellen met onvolledige informatie. Je kan een uitspraak krijgen over een situatie, bijvoorbeeld over de verdeling van een kenmerk, of over de relatie tussen meetniveau en de keuze voor een parameter van centraliteit of spreiding. Je kan gevraagd worden een variantie zelf te berekenen uit een beperkt aantal gegevens. Je bent best voorbereid als je klaar en duidelijk weet welk antwoord een statistische beschrijvende parameter kan geven. Ten tweede stel je best automatisch de vraag of de techniek je wel toelaat om je probleemstelling te beantwoorden. Is het meetniveau van de variabele die je wil bestuderen wel in overeenstemming met de keuze die je zou maken?

Leer univariate grafische voorstellingen bekijken en herinterpretieren de tabellen ook in het licht van de essentiële informatie waaraan tabellen dienen te voldoen. Dit hebben we in het inleidend hoofdstuk reeds besproken en dit wordt dus niet meer herhaald. In de lessen zal je ook geleerd worden om aan de hand van SPSS de belangrijkste univariate beschrijvende parameters te berekenen en te interpreteren.



## **Hoofdstuk 4**

### **Een inleiding in kansrekenen**

#### **1. Waarom kansrekenen voor criminologen?**

Een criminoloog moet regelmatig beslissingen nemen onder onzekere omstandigheden. Bij de beslissing om een pedofiel vrij te laten is er sprake van een zekere kans dat de dader hervalt. Om deze kans op zijn waarde te schatten, moeten we weten wat een kans is. Een criminoloog die een uitspraak wilt doen over de onveiligheidsbeleving van de Vlaming, kan moeilijk elke Vlaming bevragen. Zes miljoen Vlamingen een vragenlijst doen invullen, is net iets van het goede te veel. Criminologen zullen daarom een steekproef nemen. Wetenschappelijke steekproeven moeten representatief zijn voor de bevolking. Steekproeven worden op verschillende manieren getrokken. Het belangrijkste is dat er een toevalsmechanisme achter de steekproef schuilt. Dit betekent dat elk element een zelfde kans heeft om getrokken te worden. Ook hier spelen kansen een rol en is het essentieel dat je een absoluut basisvocabularium en basisprincipes die schuilgaan achter de kansrekening begrijpt. Een criminoloog kan in de praktijk geconfronteerd worden met de vraag van een opdrachtgever om een steekproef te nemen van het cliënteel waar hij of zij mee werkt, met de vraag een bepaalde dienst te evalueren. Welke soort steekproef dient de criminoloog dan te nemen? We zullen verderop zien dat er nogal wat mogelijkheden bestaan. Toevalssteekproeven verschillen van elkaar in de kans op een bepaalde uitkomst. Je zou bijvoorbeeld elk element in de populatie een nummer kunnen geven en dan via toevalsmechanisme A (zonder teruglegging) je steekproef bepalen, maar je zou ook elke respondent die je getrokken hebt opnieuw de kans kunnen geven om getrokken te worden. Waarom bestaan er zulke verschillen en waarom zouden we kiezen voor op maat gemaakte steekproeven? In dit hoofdstuk wordt hierop een antwoord geboden. Aan de hand van **drie kansdefinities** maken we duidelijk hoe een elementaire kans kan bepaald worden. Daarna geven we enkele kansregels om kansen te berekenen in iets ingewikkeldere situaties. Dit wordt gedaan via de introductie van het concept voorwaardelijke kans, i.e. de kans op een gebeurtenis onder een bepaalde andere conditie.

## 2. Kansdefinities

Experimenten, waarvan wordt geëist dat ze een onbeperkt aantal keren kunnen worden herhaald onder gelijkblijvende omstandigheden, worden kansexperimenten genoemd. Een steekproef trekken is ook zo een situatie. Een **kans** is steeds een getal tussen 0 en 1 (een proportie), maar vaak drukken we deze uit in een percentage. De notatie van een kans is de letter **P**. De kans om een zes te gooien met één dobbelsteen wordt genoteerd als **P(6)** en wordt uitgesproken als de kans op een 6. De vraag die we ons stellen, is: kunnen we bepaalde kansen berekenen? En zo ja, hoe doen we dat dan? Welke is de kans om te winnen in een loterij? Welke is de kans om in de gevangenis te belanden? Welke is de kans dat men slachtoffer wordt van een geweldsdelict als men in een grote stad woont?

In een aantal gevallen is dit heel eenvoudig. Je berekent de **kansdefinitie van Laplace**. De kansdefinitie van Laplace is het aantal uitkomsten waarin je geïnteresseerd bent, gedeeld door het aantal mogelijke uitkomsten van een kansexperiment. De kansdefinitie van Laplace is een theoretische of objectieve kansdefinitie.

**De formule die daar bij hoort is:  $N(A)/N$  waarbij**

A= uitkomst van het kansexperiment (de gebeurtenis) waar je in geïnteresseerd bent

$N(A)=$  het aantal uitkomsten waar je in geïnteresseerd bent

N= het aantal uitkomsten

Als je met een dobbelsteen gooit en je wilt weten hoe groot de kans is dat je een even getal gooit als je een keer mag gooien, ga je als volgt te werk:

A zijn de uitkomsten waarin je geïnteresseerd bent: 2,4,6

Het aantal uitkomsten  $N(A)$  waarin je geïnteresseerd bent is 3 (je bent in die drie gebeurtenissen geïnteresseerd). Het aantal mogelijke uitkomsten is echter 1, 2, 3, 4, 5 en 6. Er zijn dus zes mogelijkheden terwijl er slechts drie interessant zijn. De kans om even te gooien is:  $P(A)= N(A)/N$  ofwel  $3/6= 50\%$ .

Het is niet altijd mogelijk om de objectieve of theoretische kans te bepalen. Als dat onmogelijk is, kan men wel de relatieve frequentie bepalen. Deze noemen we ook vaak de experimentele kans en is gebaseerd op eigen onderzoek.

De relatieve frequentie is te bepalen door het kansexperiment te herhalen en vervolgens te kijken hoe vaak de uitkomst die je interesseert zich voordoet ten opzichte van het totaal aantal herhaalde kansexperimenten.

Hierbij hoort de formule:

$$P(A) = n(A)/n$$

Waarbij:

*n(A) = het aantal uitkomsten waarin je geïnteresseerd bent bij herhaling van het experiment*

*n = het aantal herhalingen van het experiment*

Hier komen we al vrij dicht in de buurt van wat beleidsmakers soms van criminologen willen weten. In veel “real life” situaties is een kans niet op logischerwijze te bepalen. Als je zelf een nieuwe weekendbijverdienste in een kroeg zoekt en je wilt weten hoe groot de kans is dat er ‘s weekends in de kroeg waar je overweegt te gaan werken, vechtpartijen uitbreken, dan kan je daarover cijfers verzamelen. Je kan natuurlijk niet elk weekend cijfers verzamelen. Als je maar enkele weekends observaties uitvoert, dan zal je eigen onderzoek weinig precieze uitkomsten genereren. Maar als je er een jaar lang elk weekend observaties maakt, dan weet je toch veel preciezer hoe groot de kans is dat de gebeurtenis waarin je geïnteresseerd bent, effectief kan plaatsvinden. Als je dertig weekends hebt geobserveerd en we hebben op 3 weekends problemen geobserveerd, dan is de relatieve kans (volgens jouw bevindingen) 3/30 of 1/10 of 10%. Kansen worden hier gedefinieerd in termen van veelvuldig herhalen. Hier speelt wat men noemt **de experimentele wet**: naarmate het aantal herhalingen van een toevalsproces toeneemt, zullen de kansen van de elementen van S (de steekproef) zich meer en meer stabiliseren. Bij een groot aantal herhalingen wordt een stabiele waarde beschouwd als de kans dat element uit S zich voordoet. Criminologen zien steekproefgegevens als uitkomst van een kansproces!

Het is niet altijd mogelijk om een kansexperiment uit te voeren. Wanneer je mensen op straat hoort praten over hun eigen impressies, over kansen, dan spreken we over **subjectieve kansen**, in tegenstelling tot objectieve kansen. De subjectieve kans is de eigen inschatting en

is dus gebaseerd op de perceptie van het individu. De kans dat student Jan Janssens later een succesvol gevangenisdirecteur zal worden, staat niet geschreven op het diploma van de student, maar kan de persoon zelf trachten in te schatten. Het resultaat van deze eigen inschatting is de **subjectieve kans**.

### 3. Kansregels

Om kansen te berekenen, bestaan er verschillende regels. In deze syllabus dienen volgende kansregels gekend te zijn: **de algemene somregel, de speciale somregel, de algemene productregel, de speciale productregel, de complementregel.** Deze regels zijn van belang omdat kansrekenen de wiskundige beschrijving van toevalsfenomenen is. Inzicht in de absolute basis hiervan is noodzakelijk om te begrijpen hoe we aan inductieve statistiek doen, of anders gezegd hoe we als criminologen in staat zijn om op basis van één steekproef uitspraken te doen over een onderzoekspopulatie. We gaan in deze syllabus enkel en slechts enkel de basisregels toelichten.

**De algemene somregel** luidt als volgt:

$$P(A \text{ of } B) = P(A) + P(B) - P(A \text{ en } B)$$

We zijn hier geïnteresseerd in de twee gebeurtenissen A en B en we willen weten of één van deze twee gebeurtenissen afzonderlijk plaatsvindt, maar niet gezamenlijk.

Voorbeeld uit een kaartspel (52 kaarten):

13 harten

13 ruiten

13 schoppen

13 klaveren

In iedere set van kaarten zit een vrouw. Er zijn dus 4 vrouwen. Je wilt weten hoe groot de kans is dat je of een vrouw of een hartenkaart trekt.

$$P(A) = \text{Probabiliteit harten} = 13/52$$

$$P(B) = \text{Probabiliteit vrouw} = 4/52$$

$$P(\text{harten en vrouw}) = 1/52$$

Merk op dat de hartenvrouw twee maal geteld wordt, daarom moet deze er één maal afgehaald worden. A en B hebben namelijk één zaak gemeenschappelijk.

De kans op harten of vrouw is dan:  $13/52 + 4/52 - 1/52 = 16/52$  of 30.77%

**De speciale somregel** luidt als volgt:

**$P(A \text{ of } B) = P(A) + P(B)$  waarbij geldt dat A en B niets met elkaar gemeen hebben.**

Opnieuw een voorbeeld uit een kaartspel: als je wilt weten wat de kans is op harten of ruiten, dan tel je de beide kansen op.

$$P(\text{harten}) = 13/52$$

$$P(\text{ruiten}) = 13/52$$

$$P(\text{harten of ruiten}) = 13/52 + 13/52 = 50\%.$$

Precies omdat er geen dubbeltellingen zijn, kunnen we deze beide kansen optellen.

### **De algemene productregel**

In de realiteit is er vaak overlap tussen twee gebeurtenissen. Daarom moet het begrip **voorwaardelijke kans** geïntroduceerd worden. De voorwaardelijke kans op een bepaalde uitkomst is de kans op een bepaalde uitkomst als je al gedeeltelijke informatie hebt over de uitkomst van een kansexperiment. Stel dat je met een dobbelsteen gooit en je wil weten hoe groot de kans is dat je hoogstens drie gooit. Omdat je weet dat er evenveel getallen kleiner of gelijk aan drie zijn dan groter dan drie, heb je voorkennis en kan je de voorwaardelijke kans berekenen. **De voorwaardelijke kans is de kans op gebeurtenis A, onder conditie van gebeurtenis B.** In de bivariate statistiek zullen we vaak met conditionele kansen geconfronteerd worden, bijvoorbeeld de kans dat iemand een veelpleger is, gegeven het feit dat men jongen is, versus de kans dat iemand veelpleger is, gegeven de kans dat iemand meisje is.

De notatie van een **voorwaardelijke of conditionele kans** luidt:

$$P(B | A) = P(A \text{ en } B) / P(A)$$

Dit is de kans op B gegeven A. Als we deze regel toepassen en we willen weten wat de kans is op het gooien van een getal dat niet hoger is dan drie (de kans op B), gegeven het feit dat het een even getal moet zijn (de kans op A) dan bekomen we:

$P(A \text{ en } B) =$  de probabilititeit dat een getal even is EN hoogstens drie is=  $1/6$ . Enkel het getal twee is immers kleiner of gelijk aan drie en even.

$P(A) =$  de probabilititeit dat een getal even is=  $3/6$

De voorwaardelijke kans=  $1/6$  gedeeld door  $3/6$  of  $1/6 * 6/3 = 6/18$  of  $1/3$  of  $33.33\%$ .

### **De speciale productregel**

In veel gevallen is vooraf ontvangen informatie die je krijgt bij een kansexperiment niet relevant voor het kansprobleem. Zo is de opmerking dat het in New York file rijden is op zaterdag weinig relevant om de kans te kennen dat je op zaterdag in Brussel file zal rijden. Het effect is dan dat  $P(B | A) = P(B)$ . De kans dat A en B samen voorkomen is dan  $P(B) * P(A)$ . Dit noemen we de **speciale productregel**.

We passen deze regel toe aan de hand van een voorbeeld. Stel dat we een vaas hebben met zes zwarte en vier rode ballen. We halen er toevalsgewijs een bal uit en leggen die daarna terug. Methodologen noemen dit een steekproef met teruglegging. De kans dat we de eerste keer een zwarte bal trekken, heeft in zulke situatie geen enkele invloed op de kans dat we de tweede keer een zwarte bal trekken. We geven nog een voorbeeld: we trekken een steekproef met teruglegging van eerstejaarsstudenten criminologie. We weten dat er 60% meisjes en 40% jongens zijn. Bij de eerste trekking wordt een meisje getrokken. Er is teruglegging dus hetzelfde meisje kan opnieuw geselecteerd worden. Wat is de kans dat opnieuw een meisje getrokken wordt? Door de teruglegging is de kans opnieuw 60%. De kans dat de tweede keer opnieuw een meisje geselecteerd wordt is  $60\% * 60\% = 36\%$ .

### **De complementregel**

Het kan voorkomen dat een kans lastig te berekenen is vanwege de complexiteit van de vraagstelling. In een aantal gevallen kan het daarom handig zijn om naar het tegenovergestelde probleem te kijken. Als dit probleem eenvoudiger op te lossen is, kan men aan de hand van de complementregel de kans berekenen.

$$P(A) = 1 - P(\text{complement van } A)$$

Als de kans om een zes te gooien met één dobbelsteen  $1/6$  is, dan is de kans om geen zes te gooien  $1 - (1/6) = 5/6$ . Als je weet dat de kans op het geslaagd zijn in de eerste zittijd van het eerste bachelorjaar criminologie  $1/10$  is, dan weet je tegelijk dat je  $90\%$  kans hebt om niet geslaagd te zijn in de eerste zittijd van het eerste bachelorjaar criminologie.

#### 4. Permutaties en combinaties

Bij kansrekenen moet er vaak veel geteld worden om bij het kansexperiment het aantal elementen te bepalen waar je in geïnteresseerd bent. Een hulp hierbij is het aantal permutaties en combinaties. Het aantal **permutaties** is **het aantal manieren waarop je een aantal verschillende objecten ten opzichte van elkaar** kan plaatsen. De volgorde is dus van belang. Als we onszelf de vraag stellen op hoeveel manieren we de letters A, B, C ten opzichte van elkaar kunnen plaatsen, dan zien we:

ABC, ACB, BAC, BCA, CAB, CBA

Dit zijn zes mogelijkheden. De redenering bij permutaties: voor de eerste letter zijn er drie mogelijkheden, namelijk A, B of C. Als we voor de eerste letter een keuze gemaakt hebben, dan blijven er voor de tweede letter nog 2 mogelijkheden over. De derde letter ligt dan vast, er blijft immers maar één letter meer over. **Het aantal permutaties van n objecten wordt weergegeven als  $n!$  (speak uit; n faculteit) en is:  $n*(n-1)...*1$ .**

vb:  $5!=5*4*3*2*1=120$

Bij combinaties zijn er bij een groep objecten twee subgroepen. Binnen iedere subgroep zijn de elementen niet van elkaar te onderscheiden. We tellen het aantal manieren om de objecten ten opzichte van elkaar neer te zetten.

We tellen het aantal combinaties dat er mogelijk is met de letters A, A, B, B en B.

AABBB, ABABB, ABBAB, ABBBA, BAABB, BABAB, BABBA, BBAAB, BBABA, BBBAA

We zien dat er tien combinaties zijn. Dit hadden we ook als volgt kunnen vinden: als alle letters verschillend zijn, zijn er  $5!=120$  mogelijkheden. Er zijn 2 A's. Het verwisselen van twee A's levert een dubbeltelling op. De 2 A's zijn op  $2!=2$  manieren te verwisselen. We moeten delen door deze dubbeltelling. Er zijn 3 B's. Het verwisselen van die B's levert ook dubbeltellingen op. In totaliteit kunnen we bij het verwisselen van de 3 B's  $3!=6$  dubbeltellingen vinden. We moeten ook door deze dubbeltellingen delen.

De berekening is:  $5!/(2!*3!)=120/12=10$

Het aantal combinaties van n (in het voorbeeld 5) elementen met een subgroep van k (2) gelijke elementen en een tweede subgroep van n-k ( $5-2=3$ ) gelijke elementen is gelijk aan 10.

Het aantal combinaties van  $k$  elementen uit een verzameling van  $n$  elementen wordt formeel genoteerd als de binomiaalcoëfficiënt  $\binom{n}{k}$  (*speek uit: n over k*). Dit wordt ook als volgt

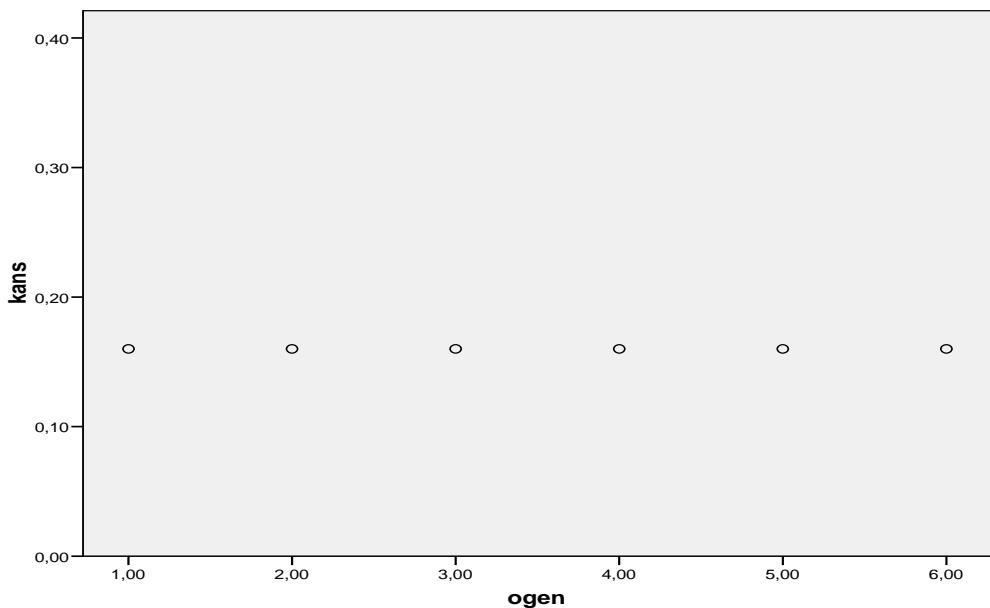
genoteerd: 
$$\binom{n}{k} = \frac{n!}{(n - k)! \cdot k!}$$
 Deze binomiaalcoëfficiënt is nodig voor het berekenen van kansen dat willekeurige gebeurtenissen voorkomen in bepaalde combinaties.

## 5. Kansvariabelen en de binomiale verdeling

Een kansvariabele geeft aan in welk getal een kansexperiment resulteert. De notatie van een kansvariabele is bijvoorbeeld  $\underline{k}$  of  $\underline{x}$ . Het streepje onder de letters verwijst naar het feit dat we te maken hebben met een kansexperiment en dat niet bij voorbaat vast staat wat de uitkomst zal zijn. Zo kunnen we kijken naar  $\underline{k}$ : het aantal ogen van een dobbelsteen bij eenmaal gooien met de uitkomsten (1, 2, 3, 4, 5, 6). In criminologisch onderzoek kijken we bijvoorbeeld naar de kans op recidive nadat 30 personen een alternatieve straf hebben gevuld. Kansen kunnen berekend worden voor **continue** en **discrete** variabelen. Het voorbeeld van de ogen van de dobbelsteen en het voorbeeld van de kans op recidive zijn voorbeelden van een discrete kansverdeling. Drie personen van de dertig kunnen recidiveren, maar dat kunnen er ook vier zijn. Tussenliggende waarden zijn onmogelijk. Als we de kansen willen berekenen van de lengte van de files in Brussel, dan kan die file 3 kilometer zijn, maar ook 3.21 of 4.26. De kansen op iedere uitkomst in een kansexperiment is weer te geven met een kansfunctie. Kijken we naar het voorbeeld van  $k$ : het aantal ogen van een dobbelsteen bij eenmaal gooien, dan zijn de kansen op de uitkomsten:

$$P(\underline{k}=1) = 1/6 ; P(\underline{k}=2) = 1/6 ; P(\underline{k}=3) = 1/6 ; P(\underline{k}=4) = 1/6 ; P(\underline{k}=5) = 1/6 ; P(\underline{k}=6) = 1/6$$

**Figuur: kansverdeling voor het gooien van ogen met één dobbelsteen**



We kunnen de uitkomsten van een kansfunctie cumuleren:

$$P(\underline{k} \leq 1) = 1/6$$

$$P(\underline{k} \leq 2) = 2/6$$

$$P(\underline{k} \leq 3) = 3/6$$

$$P(\underline{k} \leq 4) = 4/6$$

$$P(\underline{k} \leq 5) = 5/6$$

$$P(\underline{k} \leq 6) = 6/6 \text{ of } 1$$

We noemen dit overzicht de **verdelingsfunctie**. Een delingsfunctie is **het overzicht van de uitkomsten van een kansvariabele met de hierbij behorende gecumuleerde kansen**. Een kenmerk van een kansvariabele is de verwachte waarde. De verwachte waarde is de som van de uitkomsten vermenigvuldigd met de kans op iedere uitkomst.

In een formule wordt de verwachte waarde als volgt weergegeven:

$$E \underline{k} = \sum P(\underline{k}=k) * k$$

Bij het voorbeeld met de dobbelsteen is dit:

$$E \underline{k} = 1/6 * 1 + 1/6 * 2 + 1/6 * 3 + 1/6 * 4 + 1/6 * 5 + 1/6 * 6 = 3,5.$$

Dit betekent dat als je heel vaak (en in theorie oneindig vaak) met een dobbelsteen gooit en het gemiddelde berekent van je resultaten, je uiteindelijk 3,5 zal vinden als gemiddelde.

Tevens kan de standaardafwijking bij een kansvariabele bepaald worden. De standaardafwijking bij een discrete kansvariabele kan berekend worden met de formule:

$$\sigma = \sqrt{\sum P(\underline{k} = k)(k - E\underline{k})^2}$$

In het voorbeeld van de dobbelsteen is dit 1,71. Deze waarde vind je ook door heel vaak (oneindig vaak) het kansexperiment te herhalen en vervolgens de standaardafwijking ervan te berekenen.

## 6. De binomiale verdeling

In kwantitatief criminologisch onderzoek zijn kansvariabelen van groot belang. We geven opnieuw het voorbeeld van recidive-onderzoek. We willen weten hoe groot de kans is dat niemand tot iedereen die een alternatieve sanctie heeft gekregen, hervalt in oude gewoonten of recidiveert. M.a.w. we willen eigenlijk weten hoe effectief een behandelingsprogramma is. De kansvariabele die voor ons van belang is in het recidivevoorbeeld, is  $k$ : het aantal personen uit een groep van 5 personen die een alternatieve maatregel hebben gekregen, die na de alternatieve maatregel toch hervalt.

De mogelijke uitkomsten zijn:

$$P(\underline{k}=0) ; P(\underline{k}=1) ; P(\underline{k}=2) ; \dots ; P(\underline{k}=5)$$

Hoe worden deze kansen nu berekend? Hoe groot is  $P(\underline{k} = 0)$ ? Met andere woorden, hoe groot is de kans dat geen van de personen hervalt? Stel dat we uit een grootschalig onderzoek weten dat 40% van de personen die de alternatieve maatregel kregen opgelegd, hervalt. Hoe zit het dan bij onze groep van 5 personen? De kans dat iemand niet hervalt is dus 60%. Alle vijf de personen mogen in ons voorbeeld niet hervallen. We kunnen dus de speciale productregel toepassen om de kans te berekenen. Het feit dat één persoon hervalt heeft statistisch niets te maken met het feit dat een andere persoon hervalt.

De kans in onze groep is dus:  $P(k=0) = 0.40^0 * 0.60^5 = 0.07776$

Wat als we willen weten wat de kans is dat precies één persoon hervalt?:  $P(\underline{k} = 1)$ . Deze persoon benoemen we met de letter J, alle andere met N. Dit geeft ons verschillende combinatiemogelijkheden:

JNNNN	(De eerste persoon hervalt, de overigen niet)
NJNNN	(Enkel de tweede persoon hervalt)
NNJNN	(Enkel de derde persoon hervalt)
NNNIN	(Enkel de vierde persoon hervalt)
NNNNJ	(Enkel de laatste persoon hervalt)

Of anders: de formule van de binomiaalcoëfficiënt  $\binom{n}{k} = n!/(k!(n-k)!)$  toegepast op ons voorbeeld:

$$5!/(1!*4!) = 5$$

Er zijn dus 5 manieren waarop we precies één persoon kunnen vinden die hervalt.

Hoe groot is dan  $P(\text{JNNNN})$ ? M.a.w. hoe groot is de kans dat de combinatie JNNNN voorvalt? Deze kans vinden we door de toepassing van de speciale productregel:  $0.40^1 * 0.60^4 = 0.0518$ . Het voorkomen van één van de andere vier combinaties kent dezelfde kans.

Wanneer we de kans  $P(\underline{k}=1)$  willen kennen. Anders gezegd, wanneer de combinatiemogelijkheid niet van belang is, maar we willen weten hoe groot de kans is dat een willekeurig iemand recidiveert, moeten we de speciale productregel toepassen maal het aantal manieren waarop we de k (in ons voorbeeld: het aantal personen uit een groep van 5 personen die een alternatieve maatregel hebben gekregen, die na de alternatieve maatregel toch hervalt) kunnen verdelen over de n waarnemingen. M.a.w. we gebruiken de binomiaalcoëfficiënt om dit aantal te vinden.

De kans dat een willekeurig iemand recidiveert is dan:

$$P(\underline{k}=1) : (0.40^1 * 0.60^4) * 5 = 0.2592$$

Op dezelfde manier kunnen we de kans berekenen dat exact de volgende combinatie voorkomt: JJNNN (= De eerste twee personen recidiveren, de drie laatste niet).

$$\text{De kans } P(\text{JJNNN}) = 0.40^2 * 0.60^3 = 0.0346.$$

Het aantal combinaties van 2 uit 5 ('vijf over twee') is:

$$5!/(2!*3!) = 10 \text{ mogelijkheden}$$

De kans dat twee willekeurige personen recidiveren is dan:

$$P(\underline{k} = 2) : (0.40^2 * 0.60^3) * 10 = 0.3456$$

En 3...

$$P(\underline{k} = 3) : (0.40^3 * 0.60^2) * 10 = 0.2304$$

$$P(\underline{k} = 4) : (0.40^4 * 0.60^1) * 5 = 0.0768$$

$$P(\underline{k} = 5) : (0.40^5 * 0.60^0) = 0.01024$$

Als we de kansen cumuleren, krijgen we het volgende:

$$P(\underline{k} \leq 0) = P(\underline{k} = 0) = 0.07776$$

$$P(\underline{k} \leq 1) = P(\underline{k} \leq 0) + (\underline{k} = 1) = 0.07776 + 0.2592 = 0.33696$$

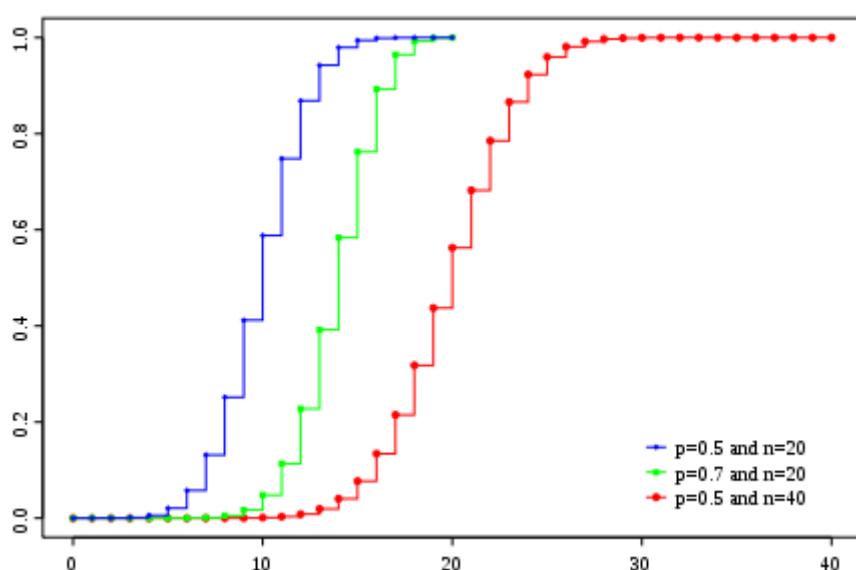
$$P(\underline{k} \leq 2) = P(\underline{k} \leq 1) + (\underline{k} = 2) = 0.33696 + 0.3456 = 0.68256$$

$$P(\underline{k} \leq 3) = P(\underline{k} \leq 2) + (\underline{k} = 3) = 0.68256 + 0.2304 = 0.91296$$

$$P(\underline{k} \leq 4) = P(\underline{k} \leq 3) + (\underline{k} = 4) = 0.91296 + 0.0768 = 0.98976$$

$$P(\underline{k} \leq 5) = P(\underline{k} \leq 4) + (\underline{k} = 5) = 0.98976 + 0.01024 = 1$$

**Figuur: Cumulatieve binomiale verdelingen**



## **7. De binomiale verdeling gaat over in een normale verdeling**

De binomiale verdeling gaat over in een normale verdeling. We geven hieronder voorbeelden. Hoe meer tentamens en hoe meer studieresultaten, hoe meer combinaties van voldoenden en onvoldoenden er mogelijk zijn. Stel dat er slechts twee mogelijke uitslagen bij een tentamen zijn: Voldoende (V) en onvoldoende (O). Stel dat de kansen als volgt verdeeld zijn: de kans op een voldoende is 60% en op een onvoldoende 40%, of:  $P(V) = 0,6$  en  $P(O) = 0,4$

Bij één tentamen zijn de volgende uitkomsten mogelijk:

$$P(V) = 0,6$$

$$P(O) = 0,4$$

Bij twee tentamens zijn de volgende uitkomsten mogelijk:

$$P(VV) = P(V) * P(V) = 0.6 * 0.6 = 0.36$$

$$P(OV) = P(O) * P(V) = 0.4 * 0.6 = 0.24$$

$$P(VO) = P(V) * P(O) = 0.6 * 0.4 = 0.24$$

$$P(OO) = P(O) * P(O) = 0.4 * 0.4 = 0.16$$

Kans op voldoende bij twee tentamens (twee 'trekkingen';  $n = 2$ ):

$$\text{Kans op 0 voldoenden} = P(OO) = 0.16$$

$$\text{Kans op 1 voldoende} = P(OV) + P(VO) = 0.24 + 0.24 = 0.48 \text{ of anders: } (0.40^1 * 0.60^1)^*2 = 0.48$$

$$\text{Kans op 2 voldoenden} = P(VV) = 0.60^2 = 0.36$$

**Kansen bij drie tentamens:**

$$P(OOO) = 0,4 * 0,4 * 0,4 = 0,064$$

$$P(OOV) = 0,4 * 0,4 * 0,6 = 0,096$$

$$P(OVO) = 0,4 * 0,6 * 0,4 = 0,096$$

$$P(VOO) = 0,6 * 0,4 * 0,4 = 0,096$$

$$P(OVV) = 0,4 * 0,6 * 0,6 = 0,144$$

$$P(VOV) = 0,6 * 0,4 * 0,6 = 0,144$$

$$P(VVO) = 0,6 * 0,6 * 0,4 = 0,144$$

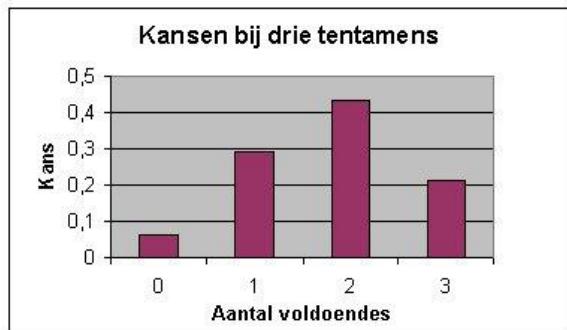
$$P(VVV) = 0,6 * 0,6 * 0,6 = 0,216$$

Kans op 0 voldoenden =  $P(000) = 0,064$

Kans op 1 voldoende =  $P(00V) + P(0VO) + P(V00) = 3 * 0,096 = 0,288$

Kans op 2 voldoenden =  $P(0VV) + P(VOV) + P(VV0) = 3 * 0,144 = 0,432$

Kans op 3 voldoenden =  $P(VVV) = 0,216$



#### Kansen bij vier tentamens ( $n = 4$ ):

$$P(VVVV) = 0,6 * 0,6 * 0,6 * 0,6 = 0,1296$$

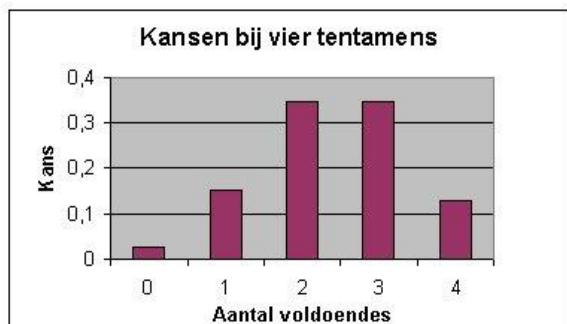
$$P(VVVO) = 0,6 * 0,6 * 0,6 * 0,4 = 0,0864$$

$$P(VVOV) = 0,6 * 0,6 * 0,4 * 0,6 = 0,0864$$

$$P(VOVV) = 0,6 * 0,4 * 0,6 * 0,6 = 0,0864$$

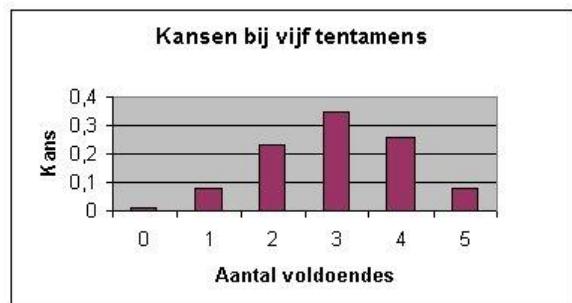
etc. tot:  $P(0000) = 0,4 * 0,4 * 0,4 * 0,4 = 0,0256$

Aantal voldoendes	Kans
0	0,0256
1	0,1536
2	0,3456
3	0,3456
4	0,1296



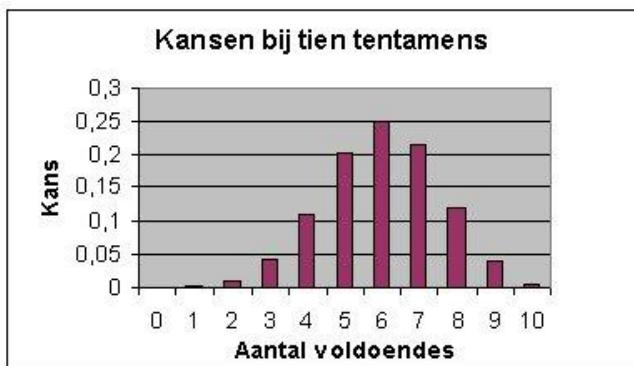
### Kansen bij vijf tentamens ( $n = 5$ ):

Aantal voldoendes	Kans
0	0,0102
1	0,0768
2	0,2304
3	0,3456
4	0,2592
5	0,0778

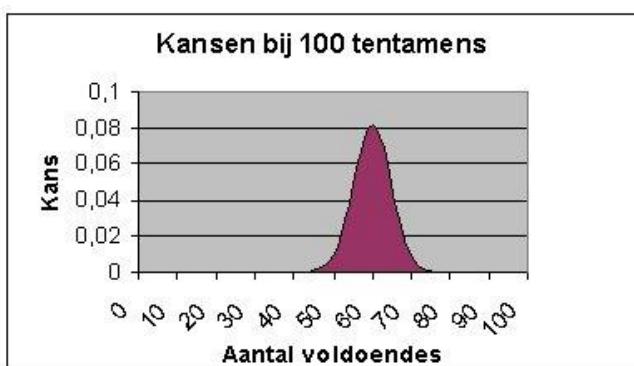


### Kansen bij tien tentamens ( $n = 10$ ):

Aantal voldoendes	Kans
0	0,000100768
1	0,001603456
2	0,010600778
3	0,0425
4	0,1115
5	0,2007
6	0,2508
7	0,2150
8	0,1209
9	0,0403
10	0,0060



**Kansen bij 100 tentamens ( $n = 100$ ):**



Wat blijkt uit deze oefening? Naarmate het aantal tentamens toeneemt, benadert de kansverdeling de normale verdeling meer en meer. De grafische afbeelding van de normale verdeling is de Gauss-curve die we in een vorig hoofdstuk besproken hebben.

## 8. Waarom is de binomiale verdeling zo belangrijk in kwantitatief criminologisch onderzoek?

Veel kenmerken waarin de criminoloog geïnteresseerd is, volgen een binomiale verdeling. Het al of niet plaatsvinden van criminale gebeurtenissen zoals slachtofferschap, is zo een voorbeeld. Deze kenmerken zijn categorische variabelen, dichotomieën, bvb 1 = slachtoffer, 0 = geen slachtoffer. Wanneer we aan statistische inferentie doen, met name het veralgemenen van steekproefresultaten naar een bredere populatie, zullen we gebruik maken van de kenmerken van deze verdeling om uitspraken te doen over de veralgemeenbaarheid van onze bevindingen. Dit komt aan bod in de volgende delen.

## **9. Leerdoelen**

Dit deel heeft tot doel de studenten een aantal elementaire kansregels aan te leren. Het begrip kansdefinitie dient goed gekend te zijn. De verschillende regels, met name de algemene somregel, de speciale somregel, de algemene productregel, de speciale productregel en de complementregel dienen gekend te zijn. Het is van belang te weten in welke situatie welke kansregel van toepassing is. Dit is belangrijk met betrekking tot het begrijpen van steekproefuitkomsten. Vervolgens werd in dit hoofdstuk aandacht besteed aan permutaties en combinaties. Deze beide begrippen dienen goed van elkaar te worden onderscheiden. Permutaties en combinaties dienen zelf te kunnen worden uitgerekend en evenals de kans op een bepaalde uitkomst.



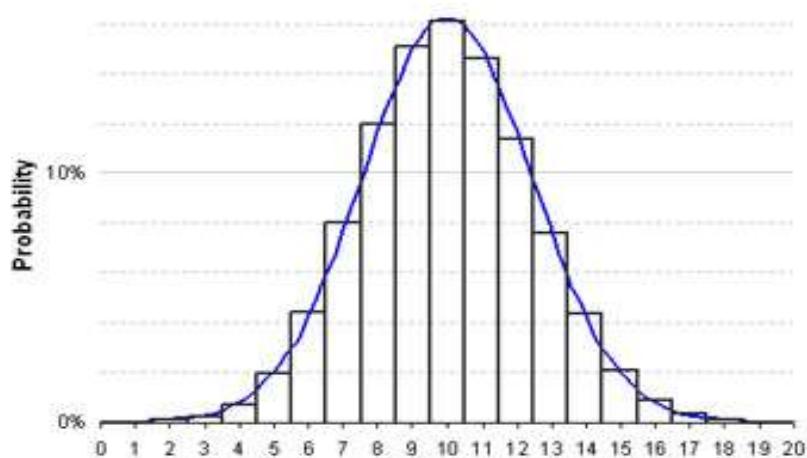
## Hoofdstuk 5

### De standaardnormale verdeling en diens eigenschappen

#### 1. Inleiding

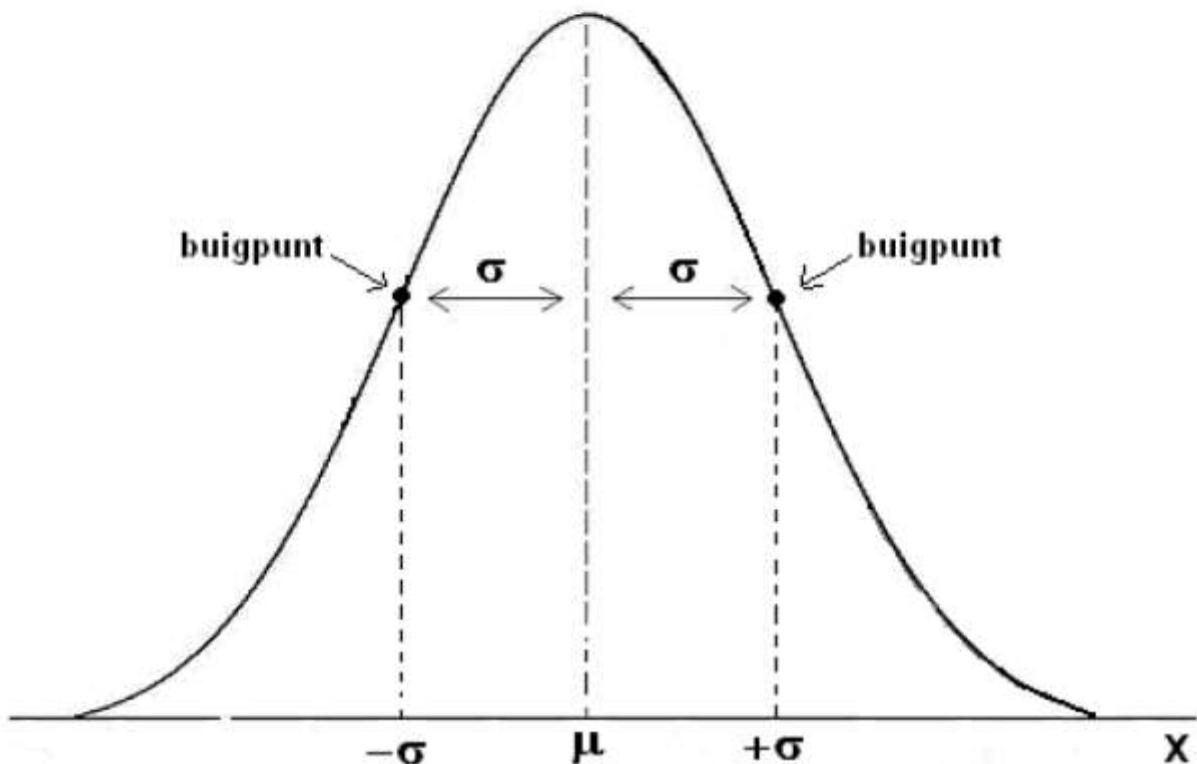
Wanneer we een continu gemeten kenmerk met haar verschillende waarden op de X-as plaatsen en op de Y-as de frequentie noteren, dan kunnen we blokjes tekenen die de frequentie weergeven. In de statistiek is men tot de vaststelling gekomen dat de aldus getekende histogrammen voor heel wat metrische kenmerken steeds een vergelijkbaar patroon hebben. Men vindt een beperkt aantal eenheden in de uiterste categorieën; de meeste waarnemingseenheden zullen eerder geconcentreerd rond het rekenkundige gemiddelde zitten. Deze verdeling noemt men ook wel **Gauss-curve of normale verdeling**. De verdeling benadert wat men noemt een **perfecte klokvorm**. Deze (theoretische) verdeling is uitermate belangrijk omdat er een aantal eigenschappen aan verbonden zijn, die we nodig hebben wanneer we later pogingen om uitspraken op basis van steekproeven te veralgemenen naar de populatie. Elke proportie van een kenmerk komt overeen met wat we **een kans** (probabiliteit) noemen. De oppervlakte onder de curve stelt de proportie 100% voor.

**Figuur: histogram met benadering van een normale verdeling**



## 2. De normale en standaardnormale verdeling

Figuur: een normale verdeling van het kenmerk “gewicht”,  $N(75,4)$



### Wat zien we in deze grafiek?

Voorerst is het belangrijk stil te staan bij de notatie die we hier gebruiken. De notatie  $N(75,4)$  wijst op een verdeling met rekenkundig gemiddelde 75 en standaardafwijking 4.

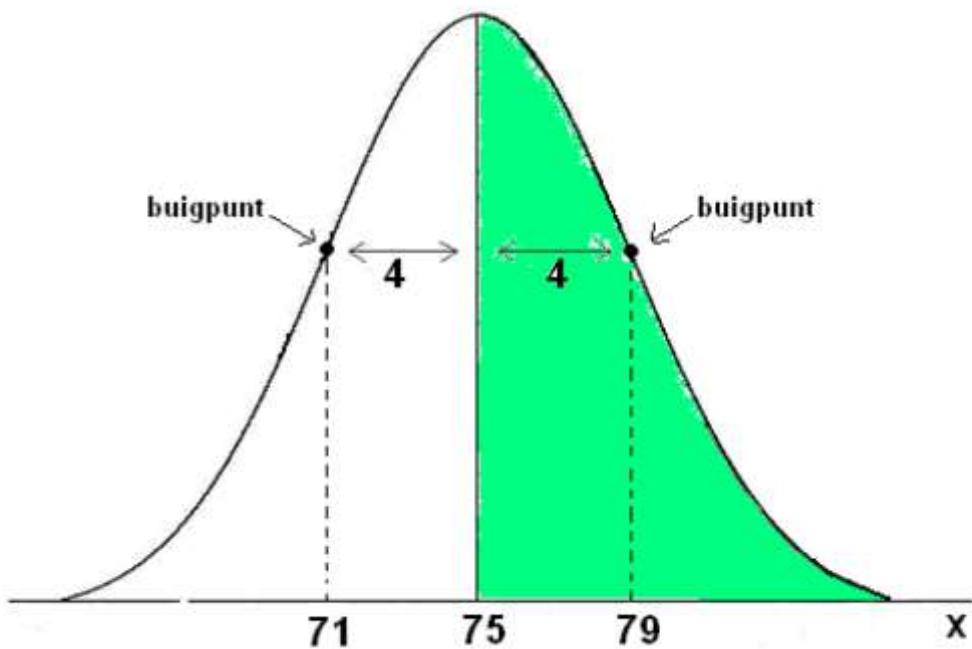
- De normale verdeling is gecentreerd om parameter  $\mu$  of “mu”, genoemd naar de Griekse letter. Deze parameter geeft de symmetrieas van de grafiek aan. Dit is het rekenkundig gemiddelde in een onderzoeks populatie.
- De grafiek heeft twee **buigpunten**, dat zijn punten waar de vorm van de kromme overgaat van ‘bol’ naar ‘hol’. De tweede parameter van de normale verdeling  $\sigma$  of “sigma” bepaalt de ligging van deze buigpunten.

**Uit wat hierboven werd gezegd trekken we een belangrijke conclusie:** als we van een normale verdeling de twee parameters  $\mu$  en  $\sigma$  kennen, kunnen we de grafiek van die normale verdeling tekenen. De hoogte van de grafiek is niet van belang zoals we later zullen zien.

In de normale verdeling (en in iedere continue verdeling) kunnen we een frequentie interpreteren als een oppervlakte onder de grafiek van de verdeling. Deze frequenties zijn te

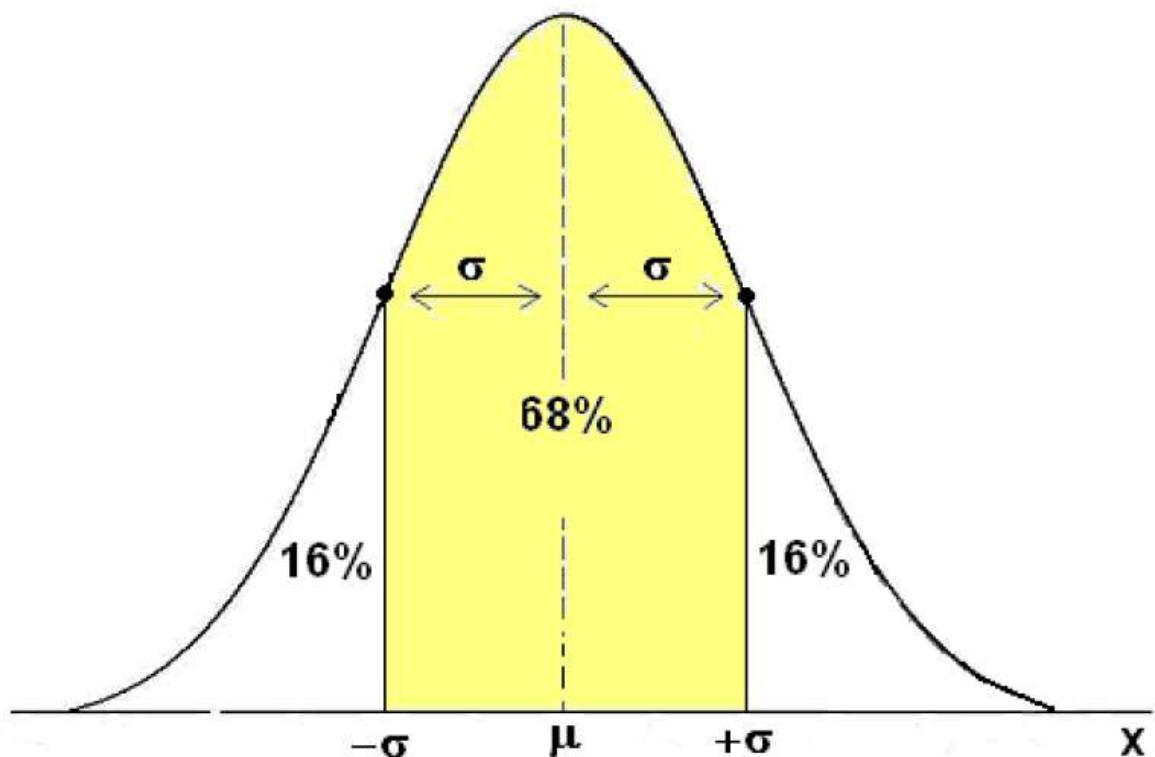
interpretieren als kansen, waarover specifiek meer in een volgend hoofdstuk. In een normale verdeling ligt 50% van de waarnemingen rechts van parameter  $\mu$ , wat eenvoudig is te begrijpen omdat de normale verdeling symmetrisch rond  $\mu$  is.

Stel bijvoorbeeld dat we weten dat in een populatie van 20-jarige Nederlandse jongens het gewicht normaal verdeeld is met een gemiddelde waarde van 75kg en een standaardafwijking van 4kg. De gewichten zijn hier op te vatten als numerieke waarden van de variabele ‘gewicht’, dus als scores. De vraag hoe groot het percentage jongeren is met een gewicht van meer dan 75kg is eenvoudig te beantwoorden. We tekenen eerst de grafiek van  $N(75,4)$  en geven in de grafiek het oppervlak rechts van  $\mu = 75$  aan. Rechts van het gemiddelde  $\mu = 75$  kg ligt 50% van het oppervlak onder de grafiek. Dus is 50% van de jongeren zwaarder dan 75kg. Als we willekeurig (“at random”) één jongere kiezen uit onze populatie is de kans dat hij/zij zwaarder is dan 75kg gelijk aan 0,5.

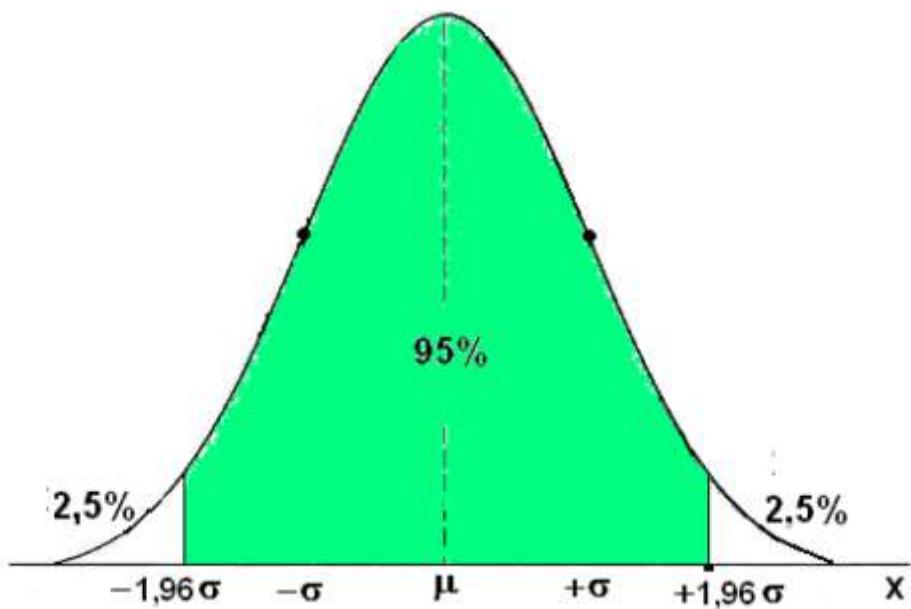


Om een schatting te kunnen maken van de proporties (= kansen) die overeenkomen met bepaalde waarden bij normaal verdeelde scores zijn de volgende drie eigenschappen van de normale verdeling van belang:

1. Bij een Normale Verdeling ligt ongeveer 68% van de scores tussen de benedengrens  $\mu - 1\sigma$  en de bovengrens  $\mu + 1\sigma$ , dus in het interval  $[\mu - 1\sigma ; \mu + 1\sigma]$ .



2. Bij een Normale Verdeling ligt 95% van de scores in het interval  $[\mu - 1,96 \sigma ; \mu + 1,96 \sigma]$ .



3. Bij een Normale Verdeling ligt 99,7% van de scores in het interval  $[\mu - 3 \sigma ; \mu + 3 \sigma]$ . De twee “staarten” bevatten dan elk 0,15% van de scores. De resolutie van het gebruikte

tekenprogramma is te grof om dit in een figuur aan te geven. Het is duidelijk dat het zéér uitzonderlijk is dat een normaal verdeelde score meer dan 3 standaardafwijkingen verwijderd is van het gemiddelde. Dit wordt ook de ‘**68-95-99**’ regel genoemd. Deze regel is cruciaal.

### 3. Van normale verdeling naar standaardnormale verdeling

Elke normale verdeling is volledig symmetrisch en *unimodaal*, zodat gemiddelde, modus en mediaan samenvallen. Hieruit volgt dat 50 procent van de waarden boven, en 50 procent van de waarden onder het gemiddelde ligt. Voor elke normale verdeling geldt dat een vast percentage van de waarden ligt tussen het gemiddelde en een bepaald getal. Echter, **normale verdelingen verschillen in termen van gemiddelde en standaardafwijking**. Om te vermijden dat men telkens de oppervlakte onder de curve voor elke verdeling moet berekenen op basis van de formule, hebben statistici gezocht naar een **standaardisering**. Hierdoor diende men de oppervlakte eenmalig te berekenen en vast te leggen in een tabel. Door te standaardiseren wordt het gemiddelde gelijkgesteld aan 0 en de standaardafwijking aan 1. Deze berekeningen houden een verandering van schaal in. De standaardnormale verdeling (ook wel *z-verdeling* genoemd) is een bijzonder geval van de normale verdeling. De standaardnormale verdeling neutraliseert de zuiver numerieke verschillen tussen normale verdelingen en geeft een algemeen overzicht van de kansverdeling, onafhankelijk van de grootte van de waarden.

De standaardnormale **z-verdeling** wordt samenvattend als volgt gekarakteriseerd: (a) het gemiddelde van de standaardnormale verdeling wordt op nul gesteld door van iedere waarde het gemiddelde van de oorspronkelijke reeks af te trekken en (b) de standaardafwijking wordt op 1 gesteld door de absolute waarde van het bij (a) berekende verschil te delen door de standaardafwijking van de oorspronkelijke reeks. Dit wordt aangetoond in de volgende paragraaf.

### 4. Z-scores en het gebruik van de tabel van de standaardnormale verdeling

We kunnen iedere normale verdeling  $N(\mu, \sigma)$  transformeren tot de zogenaamde ‘standaardnormale verdeling’, symbolisch weergegeven als ‘ $N(0,1)$ ’. Om dit te doen moeten we iedere x-score omzetten in een z-score. Hoe doen we dat nu? Een voorbeeld brengt verduidelijking. We onderzoeken de lengte van een aantal mensen. We weten dat de gemiddelde lengte 168cm is en de standaardafwijking 12cm bedraagt. We willen weten hoe groot de proportie (of het percentage) is van **de onderzoekseenheden die kleiner of gelijk**

aan 143 cm lang zijn (en dus in het hieronder gearceerde gedeelte vallen). Een z-score wordt als volgt berekend:

$$z = \frac{x - \mu}{\sigma}$$

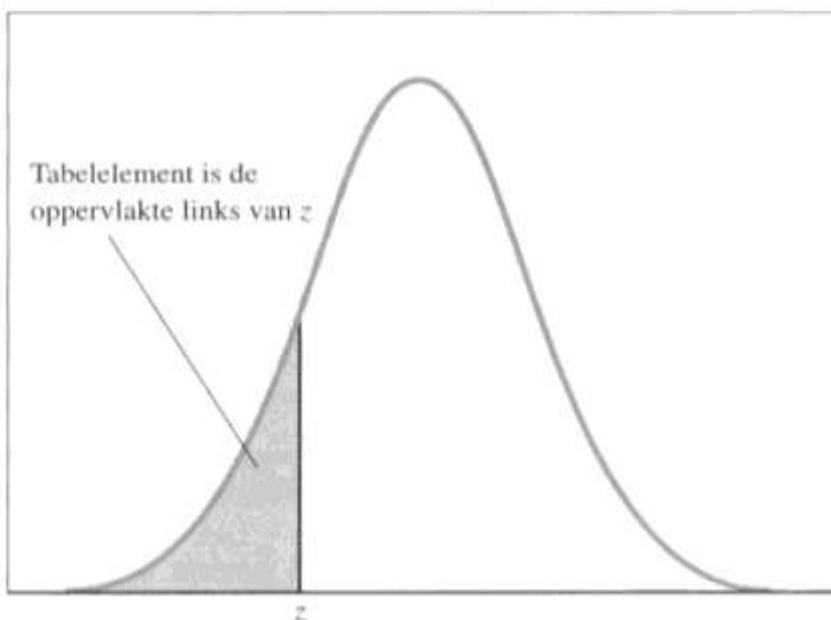
### Voorbeeld 1

De waarde van Z voor iemand die 143 cm lang is, wordt als volgt berekend:

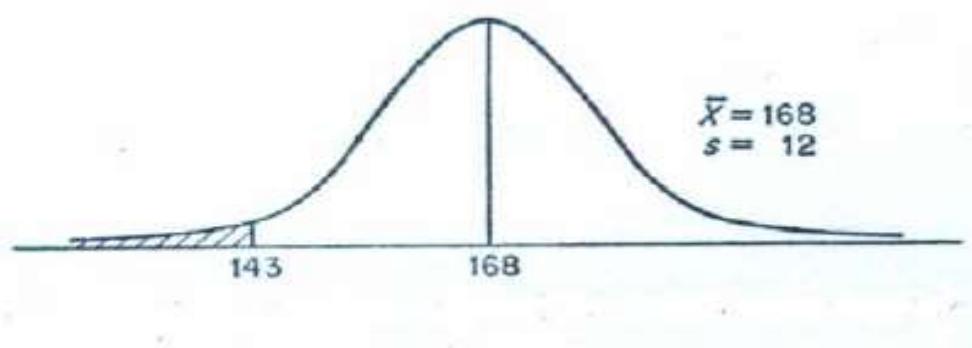
$$Z = \frac{143 - 168}{12} = \frac{-25}{12} = -2,08$$

**Wat betekent nu deze waarde van -2.08?** De oorspronkelijke ongestandaardiseerde waarde 143 ligt iets verder dan 2 standaardafwijkingen van het gemiddelde. Het negatieve teken van de Z-waarde wijst op het feit dat het gearceerde stuk links van het gemiddelde ligt. We kunnen in principe elke waarde van een metrische variabele standaardiseren. Dit betekent dat we voor elke onderzoekseenheden z-scores berekenen op de variabelen. **Voor het aflezen van de proportie van waarnemingen die in een bepaalde zone onder de normale valt, moeten we eerst de proportie van waarnemingen berekenen die overeenkomt met een score van 2.08. Dat is 98.12%. Aangezien de normale curve perfect symmetrisch is, weten we dat de proportie die overeenkomt met een z-score van -2.08 gelijk is aan 100% - 98.12% en dat is 1.88 %.** (Uit de tabel in bijlage kunnen de z-scores worden opgezocht).

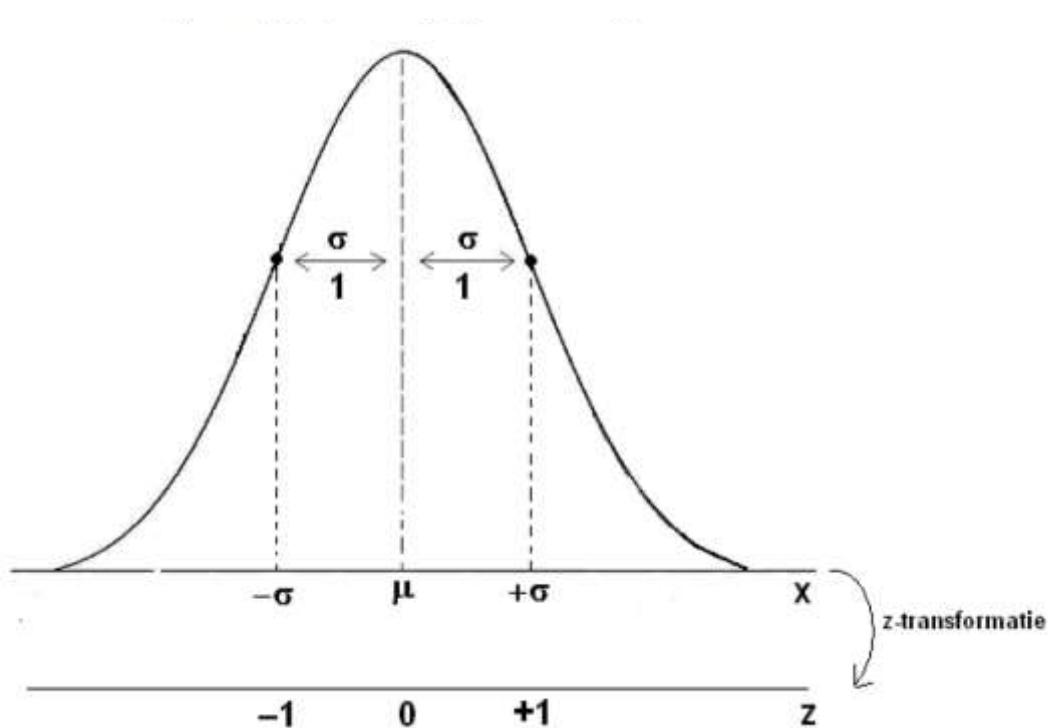
Figuur: opzoeken van tabelelementen



**Figuur: voorbeeld bij het bepalen van een z-score**

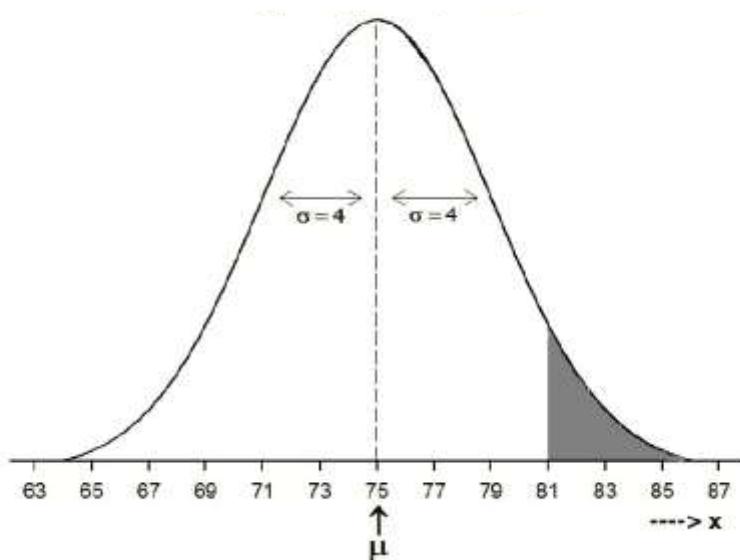


**Samenvattend:** Als de scores  $x$  normaal verdeeld zijn als  $N(\mu, \sigma)$  dan zijn de hieruit berekende z-scores normaal verdeeld als  $N(0,1)$ . Deze bijzondere normale verdeling noemen we de standaardnormale verdeling. Grafisch kunnen we het zo voorstellen dat je iedere normale verdeling kunt omzetten in een standaardnormale verdeling door de horizontale as te verschuiven ("centreren rond nul") en vervolgens uit te rekken of in te krimpen (standaarddeviatie van één). De volgende figuur geeft dit aan.



### Voorbeeld 2

In een populatie is de variabele ‘gewicht’ normaal verdeeld met gemiddelde 75kg en standaardafwijking 4kg (symbolisch:  $x \sim N(75,4)$ ). We trekken willekeurig (‘at random’) één persoon uit deze populatie. Wat is de kans dat deze persoon meer dan 81 kg weegt? Als we deze kans opvatten als een frequentie is de vraag: hoe groot is het oppervlak onder de grafiek van  $N(75,4)$  rechts van  $x = 81$ ?



$Z = (81 - 75) / 4 = 1.5$ . Als we deze kans opzoeken in de tabel van de standaardnormale verdeling, moeten we eerst zoeken welke waarde overeenkomt met 1.5 en dat is 0.9332. De kans om in deze populatie bij toeval iemand te trekken die lichter is dan 81 kg is dus 93.32%. De kans dat iemand zwaarder is dan 81 kg is dus  $1 - 0.9332$  of 0.0668 of 6.68%.

### Voorbeeld 3

Gegeven een normaalverdeelde zwangerschapsduur van gemiddeld 280 dagen met een standaardafwijking van 10 dagen. Of anders genoteerd:  $N(280,10)$ . Hoe groot is de kans dat de geboorte meer dan veertien dagen te laat plaatsvindt? We berekenen eerst de z-score die bij de vraag hoort:  $(294-280)/10 = 1,4$

Bij  $z = 1,4$  hoort een kans van 91.92 %. Dit is de kans op een geboorte voor dag 294. De kans op een geboorte die méér dan 14 dagen te laat plaatsvindt, is dus  $100 - 91.92 = 8.08\%$

#### **Voorbeeld 4**

We vertrekken opnieuw van een normaal verdeelde zwangerschapsduur van gemiddeld 280 dagen met een standaardafwijking van 10 dagen. Hoe groot is de kans dat de geboorte plaatsvindt in de periode die ligt tussen 8 dagen voor en 8 dagen na de gemiddelde datum?

Dit voorbeeld lijkt moeilijker, maar dat is het eigenlijk niet. We moeten eerst weten welke de proportie is die overeenkomt met 288 dagen. Dit is 8 dagen na het gemiddelde. De z-score die hiermee overeenkomt is 0.8 want  $(288-280)/10 = 0.8$ . De proportie die hiermee overeenkomt is 0.7881. Vervolgens moeten we weten welke de proportie is die overeenkomt met 272 dagen. Dit is 8 dagen voor het gemiddelde. De z-score die hiermee overeenkomt is -0.8, want  $(272-280)/10 = -0.8$ . De proportie die hiermee overeenkomt is  $1 - 0.7881 = 0.2119$ . De vraag behelst eigenlijk de proportie die tussen beide ligt: dus deze proportie bedraagt  $0.7881 - 0.2119$  en dat is 0.5762 of 57.62% wordt geboren tussen beide tijdstippen.

De normale verdeling (en andere delingen zoals de *binomiale verdeling* en de *chi-kwadraat verdeling* die we niet in detail bespreken in dit hoofdstuk) zijn belangrijk wanneer we uitspraken gebaseerd op één steekproef als onderdeel van een onderzoekspopulatie willen veralgemenen naar de totale onderzoekspopulatie. Hiertoe wordt gebruik gemaakt van de principes van de kansrekening en de inferentiële statistiek, die in volgende hoofdstukken worden uiteengezet. Kenmerken van statistische delingen worden dus pas echt belangrijk wanneer schattingen worden gemaakt en wanneer specifieke theorieën worden getoetst.

### **5. Leerdoelen**

Dit deel beoogt studenten de centrale eigenschappen van normale delingen bij te brengen. Studenten dienen inzicht te hebben in deze eigenschappen en deze te kunnen toepassen. Deze toepassing houdt in dat men in staat is om vraagstukken op te lossen die gerelateerd zijn aan de berekening van oppervlaktes onder de curve van de normale verdeling. Het is belangrijk in te zien dat de normale verdeling een kansverdeling is die zegt hoeveel procent van de waarnemingen een bepaalde score hebben op een bepaald kenmerk. Het principe van de transformatie dient actief te kunnen worden toegepast. Dit hoofdstuk is belangrijk in functie van het begrijpen van de principes van de inferentiële statistiek, die in latere delen aan bod komen.



## Hoofdstuk 6

### Inleiding tot de bivariate beschrijvende statistiek

#### 1. Inleiding: causale relaties versus statistische relaties

Hoewel het begrip causaliteit vaak in onze theorievorming voorkomt, kunnen wij **nooit** met statistiek het bestaan van causale relaties bevestigen. We onderzoeken enkel statistische relaties. Statistische relaties zijn nooit waterdichte bewijzen voor causale relaties. Het enige dat we als onderzoekers kunnen doen, is de kenmerken (sterkte, richting en vorm) van een verband nagaan. Causaliteit is een filosofische en theoretische kwestie. De empirische realiteit laat zich dus wel ten dele onderzoeken, op voorwaarde dat we over adequate (geldige) en betrouwbare meetinstrumenten beschikken, maar statistiek biedt nooit waterdichte bewijzen voor causaliteit. Het omgekeerde is echter wel waar: het bestaan van causaliteit impliceert statistische samenhang. En daar zit het belang van statistisch onderzoek naar de relatie tussen variabelen. Omdat David Hume heel overtuigend had beargumenteerd dat een oorzaak-gevolg relatie niet kan worden vastgesteld met het blote oog, hield hij een pleidooi voor het beschouwen van causaliteit als regelmatige samenhang. Door het herhaaldelijk samen observeren van X en Y, zo beargumenteerde hij, besluiten we tot causaliteit. Causaliteit is in deze visie dus een gewoonte van de geest en niets meer dan dat. Enkele prominente statistici, waaronder Karl Pearson, de uitvinder van de correlatiecoëfficiënt, geloofde dat causaliteit weinig belang had en dat wetenschap evengoed kon groeien door kennis te nemen van samenhangen.

Tegenover deze Humeaanse visie staat de visie dat causaliteit **reëel** is. Vroeger dacht men dat causaliteit te maken had met onvermijdelijkheid, met complete willoosheid en met onvoorkomelijkheid. Denk maar aan het deterministische gedachtengoed uit de 19<sup>de</sup> eeuw, dat kenmerkend was voor de positivistische school. Gaandeweg is men dat **determinisme** gaan omruilen voor **probabilisme**. Immers, de realiteit van het sociaal-wetenschappelijk onderzoek laat zien dat causaal onderzoek altijd leidt tot frequentieverdelingen en samenhangen die nooit perfect zijn. Maar moeten wij perfecte verbanden veronderstellen? Natuurlijk niet. De mens maakt keuzes, i.e. de mens delibereert, maar laat zich bij het maken van keuzes wel leiden door een aantal attitudes, preferenties en dergelijke meer. Keuzes worden ook gemaakt in een bepaalde context. De combinatie hiervan (desires, beliefs, opportunities, constraints) bepalen allemaal of een handeling gesteld wordt. Een causale factor is een factor die een effect teweeg brengt. Een belangrijk onderscheid is het onderscheid tussen:

- **Voldoende en noodzakelijke oorzaken:** een factor is een voldoende en noodzakelijke causale factor als het effect wordt teweeggebracht als de factor aanwezig is en het effect niet wordt teweeggebracht als de factor niet aanwezig is. Het criterium van de noodzakelijkheid in combinatie met het voldoende criterium is zeer streng. Zulke factoren vinden we niet in de criminologie.
- **Onvoldoende en noodzakelijke oorzaken:** zulke factoren brengen het effect enkel teweeg in combinatie met een reeks van factoren. Dergelijke factoren vormen het noodzakelijke geheel binnen een geheel van factoren die enkel in combinatie een effect teweegbrengen.
- **Voldoende en niet-noodzakelijke oorzaken:** deze factoren zijn zelfstandig in staat om het effect teweeg te brengen bij aanwezigheid, maar er zijn ook nog andere factoren die hetzelfde effect kunnen teweegbrengen. In tegenstelling tot de voldoende en noodzakelijke voorwaarden, is het bij de voldoende voorwaarden zo dat de factor niet noodzakelijk is. Voldoende factoren zijn veel realistischer in criminologisch onderzoek: er zijn in de geschiedenis van de criminologie veel kandidaat-voldoende factoren naar voor geschoven. Ze zijn echter op zich niet noodzakelijk, want andere factoren kunnen hetzelfde teweegbrengen.
- **Contributieve oorzaken of niet-voldoende en niet noodzakelijke oorzaken:** dit komt dichtst bij de realiteit: het gaat om factoren waarvan we weten dat ze een bijdrage leveren (een verhoging of verlaging van probabiliteiten). In het kader van multicausale problemen, zoals criminaliteitsproblemen, mogen ze nooit alleen bestudeerd worden, maar steeds in interactie met andere factoren.

In de criminologie is er sprake van **contributieve** oorzaken. Regelovertredend gedrag is een **multicausaal fenomeen waarbij verklarende factoren bestaan op meerdere niveaus (het individu met zijn persoonlijke eigenschappen en de omgeving met diens eigenschappen)**: er is niet zo maar één oorzaak van regelovertredend gedrag. In tegendeel, vaak is het zo dat er een complex samenspel is tussen causale factoren.

Bovendien moet men een onderscheid maken tussen **singulaire causaliteit** en **veralgemeenbare causaliteit**.

- **Singulaire causatie (token causation)**: causatie op het individuele niveau: waarom heeft Jon Jonsson op 27 februari 2013 zijn vrouw doodgeschoten? Het gaat hier om een individuele case studie.
- **Algemene causatie (type causation)**: waarom observeren we in een bepaalde groep (een steekproef van jongeren bijvoorbeeld) dat individuen met bepaalde kenmerken vaker gewelddadige delicten plegen?

In het meeste criminologisch verklarend onderzoek gebruiken we de statistiek om veralgemeenbare uitspraken te toetsen. We zijn meer geïnteresseerd in veralgemeenbare theorieën dan in singulaire uitspraken. Dit is begrijpelijk: veralgemeenbare theorieën geven richting aan het beleid. We willen niet enkel voorkomen dat Jan Janssens in de toekomst criminaliteit pleegt, maar we willen er vooral voor zorgen dat individuen met bepaalde kenmerken geen regelovertredend gedrag meer stellen door in te werken op die causale factoren die de sterkste samenhang vertonen in populaties. Let wel: de samenhang moet inhoudelijk en theoretisch motiveerbaar zijn. Dat is de rol van theorie en dat durven praktiserende criminologen wel eens vergeten.

## 2. Causaliteit op een bierviltje

- Causaliteit als **productie: generatieve causaliteit**: X is een productieve oorzaak van Y als X er in slaagt Y teweeg te brengen. De vraag is alleen hoe. De aanhangers van productieve causaliteitstheorieën leggen de nadruk op de rol van twee kenmerken: (1) **mechanismen** die verklaren op welke wijze de productieve causaliteit plaatsvindt en (2) een zeker “**wetmatig karakter**” (in het Engels: **lawfulness**). Volgens deze invloedrijke theorie brengt een oorzakelijke factor het gevolg teweeg via het in gang zetten van een mechanisme. In de sociale wetenschappen bestaan verschillende mechanismen, sociale mechanismen bestaan in sociale systemen en biosociale mechanismen en psychosociale mechanismen bestaan in individuen (die biosociale systemen zijn). Er kan geen sprake zijn van een mechanisme als een oorzakelijke factor de ene keer iets in gang zet en de andere keer iets anders in gang zet. Het is nu eenmaal eigen aan mechanismen dat er een zekere regelmaat is te bespeuren. Het mag niet gaan om een eenmalige samenhang. Dat “wetmatige” is vaak verkeerd begrepen geweest. Men bedoelt hier zeker geen unicausale deterministische wetmatigheden mee

zoals in de klassieke mechanica. Zo werken wetmatigheden niet in de sociale wetenschappen. Het bestaan van statistische *empirische generalisaties suggereert sociale wetmatigheden*.

- **Causaliteit als tegenfeitelijkheid.** Een causale factor brengt een gevolg teweeg, *dat niet zou zijn teweeggebracht, indien er geen interventie of manipulatie had plaatsgevonden*. Deze visie is zeer populair onder experimentele criminologen. *Causaliteit moet een verschil maken*. De belangrijkste vertegenwoordiger van de tegenfeitelijke causaliteitstheorie is James Woodward. Tegenfeitelijke causaliteit is interessant met betrekking tot de reflectie over criminaliteitspreventie: hadden we criminaliteit kunnen voorkomen indien bepaalde zaken anders waren gelopen? We kunnen door de manipulatie van de oorzaak, het gevolg teweegbrengen. Onder experimentele criminologen is deze tegenfeitelijke causaliteitstheorie van Woodward zeer populair. Opgelet: soms weten we dat we door een bepaalde manipulatie iets teveeg brengen, maar verstaan we het mechanisme niet.
- **Causaliteit als robuuste statistische afhankelijkheid:** Statistici hebben in de jaren 1960 als reactie op beschuldigingen van determinisme causaliteit gedefinieerd als robuuste causale afhankelijkheid, maar robuust diende geïnterpreteerd te worden in termen van probabilistisch. Statistisch onderzoek leidt altijd tot frequentieverdelingen en stochastische relaties.

**In de statistiek spelen we op veilig en gaan we enkel iets zeggen over robuuste afhankelijkheden. De robuuste afhankelijkheidstheorie** zegt vereenvoudigd:

- X is een oorzaak van Y als en slechts als X bestaat (de probabiliteit van X is groter dan nul).
- X is een oorzaak van Y als X temporeel eerder komt dan Y (het tijdsmoment T1 komt voor T2), er een statistische relatie is tussen X en Y: de probabiliteit van Y gegeven X is groter dan de probabiliteit van Y “tout court”. Anders gezegd: de conditionele probabiliteit is groter dan de marginale probabiliteit.
- X is een oorzaak van Y als er geen schijnverband is (“no spurious relation”: de statistische relatie tussen X en Y verdwijnt niet als er gecontroleerd wordt voor storende controlevariabelen ook wel confounders genoemd, ook wel de “ceteris paribus” conditie genoemd).

Probabilistische causaliteitstheorieën leggen het robuuste statistische asymmetrische verband uit, niet het causale verband. Maar ze hebben een belangrijke troef gehad: ze hebben de angst voor determinisme van kwantitatief onderzoek verdreven en de weg geruimd voor causaal onderzoek.

Laat ons nu een voorbeeld geven van de verschillende inhoudelijke interpretaties die aan causale verbanden gegeven worden. Inhoudelijk is het belangrijk de mechanismen van de samenleving te begrijpen, maar daartoe is statistiek maar een hulpmiddel.

- An slaagt voor het examen statistiek omdat ze een vrouw is. Vrouwen hebben een statistisch hogere kans om te slagen voor statistiek dan mannen (robust verband geslacht-studieresultaten).
- An slaagt voor het examen statistiek omdat ze ervoor gestudeerd heeft (studeren=mechanisme)
- An slaagt voor het examen statistiek omdat ze door een mentor werd getraind (trainen=interventie, die op zich inzichten verbetert).

### 3. Symmetrische en asymmetrische relaties tussen variabelen

Theoretisch gezien moeten we in de bivariate statistiek een onderscheid maken tussen **asymmetrische** relaties en **symmetrische** relaties. **Dit onderscheid is zeer belangrijk omdat het de keuze van de associatiematen zal bepalen.** Bij een **asymmetrische** statistische relatie wordt inhoudelijk verondersteld dat de ene variabele een “*causale*” *invloed* uitoefent op de andere. Een asymmetrische relatie is een dependente relatie. Er is dependentie of afhankelijkheid: Y is afhankelijk van X betekent zoveel als Y wordt mee beïnvloed door X. Nog anders gesteld: **X is een oorzaak van Y**. Een criminologisch voorbeeld: het hebben van slechte vrienden leidt tot het plegen van meer criminale handelingen. Bij **asymmetrische** relaties is er steeds een **onafhankelijke** variabele en een **afhankelijke** variabele. De afhankelijke variabele is steeds afhankelijk van de onafhankelijke variabele. In het voorbeeld is het plegen van criminale handelingen de afhankelijke variabele en het hebben van slechte vrienden de onafhankelijke variabele. **De afhankelijke variabele wordt steeds voorgesteld met de hoofdletter Y en de onafhankelijke variabele wordt steeds voorgesteld met de hoofdletter X.** We kunnen bijkomend een onderscheid maken tussen een **rechtstreeks** (direct) en een **onrechtstreeks** (indirect) effect. Een rechtstreeks effect betekent dat een onafhankelijke variabele een rechtstreekse invloed heeft op de afhankelijke variabele. Een onrechtstreeks effect betekent dat een veranderlijke X geen rechtstreeks effect heeft op de afhankelijke variabele Y, maar onrechtstreeks, via de invloed op een andere variabele Z.



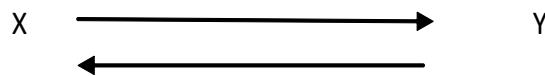
Zo heeft de sociale band van jongeren met diens ouders (variabele X) geen rechtstreeks statistisch effect op de betrokkenheid van jongeren bij crimineel gedrag (variabele Y), maar onrechtstreeks via de invloed van de ouder-kind relatie op de bereidheid van de jongere conformen normen te accepteren (variabele Z). Deze laatste variabele is doorslaggevend en dus het rechtstreekse of directe effect op de afhankelijke variabele Y.

Bij een **symmetrische** relatie kan men op theoretische gronden geen onderscheid maken tussen de beide variabelen. Het enige dat we kunnen zeggen is dat de beide kenmerken *samenhangen of correleren*. Er bestaat een verband, maar we beschikken niet over geldige argumenten om te zeggen hoe de invloed verloopt.

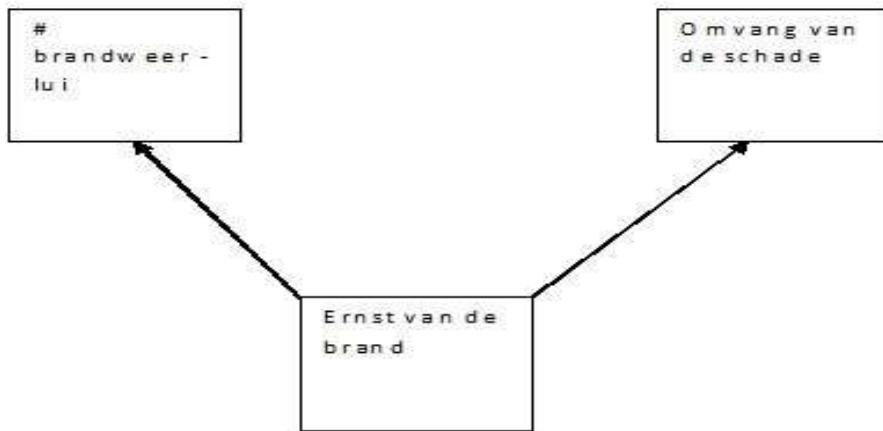
Er zijn 3 theoretische situaties te onderscheiden waarin kan gesproken worden van een symmetrische relatie tussen de variabelen X en Y.



Ten eerste kan er sprake zijn van een **wederzijdse (reciproque) invloed** tussen beide variabelen. In dat geval is er sprake van invloed van X op Y, maar ook van Y op X. Een bekend voorbeeld is de relatie tussen studiemotivatie en prestatie, een bekend criminologisch voorbeeld is het verband tussen de ruimtelijke concentratie van overlast en de ruimtelijke concentratie van criminaliteit op straat. Het kan dat overlast aanzet tot het plegen van criminaliteit, maar het kan dat criminaliteit in de straat op zijn beurt verdere overlast in de hand werkt.



Ten tweede kan er sprake zijn van een **schijnverband**. In dat laatste geval is er sprake van een **gemeenschappelijke oorzaak**.



In bovenstaand voorbeeld stelt men vast dat hoe meer brandweerlui aanwezig zijn op de plaats van de brand, hoe groter de schade is aan het gebouw. Moet men daaruit afleiden dat grote aantallen brandweerlui meer schade veroorzaken dan kleinere aantallen? Nee! Er is een gemeenschappelijke oorzaak die het schijnverband verklaart. Hoe ernstiger de brand, hoe meer brandweerlui opgetrommeld worden, en hoe ernstiger de brand is, hoe groter de schade ook is.

Ten derde kan er een samenhang bestaan tussen twee variabelen omdat deze **indicatoren zijn van hetzelfde begrip**. Zo kan er een sterke samenhang bestaan tussen het percentage buurtbewoners dat slachtoffer wordt van woninginbraak en de politiestatistieken voor woninginbraak, omdat deze beide metingen zijn van hetzelfde begrip. In dat geval vertonen beide metingen meetfouten en is de doorsnede het gevolg van het gemeenschappelijke achterliggende theoretische kenmerk dat beide indicatoren delen.

#### 4. Doelstelling van de bivariate beschrijvende statistiek

Indien we de relatie tussen twee kenmerken willen beschrijven, maken we gebruik van de *bivariate beschrijvende statistiek*. Bivariate analyses beschrijven de relatie tussen *twee variabelen*. Afhankelijk van het meetniveau wordt een beschrijvende bivariate analysetechniek gekozen. We besteden in dit handboek zowel aandacht aan *associatiematen* voor *categorische* of *non-parametrische* variabelen als aan associatiematen voor en *metrische* variabelen.

Alhoewel we soms enkel in de frequentieverdeling van één variabele zijn geïnteresseerd, gaat onze aandacht in de criminologie vaker uit naar de relatie tussen twee of meer variabelen.

Bestaat er een verband tussen twee variabelen? Zo ja, dan willen criminologen volgende zaken weten:

- **(1) hoe sterk is deze relatie;**
- **(2) welke richting (+ of -) neemt deze relatie aan en;**
- **(3) welke vorm (lineair of curvilineair) neemt deze relatie aan?**

De doelstelling van de bivariate beschrijvende statistiek bestaat er in de belangrijkste aspecten van een *relatie tussen twee variabelen* samen te vatten op een eenduidige en economische manier. Wanneer we de waarden van de elementen in onze steekproef op twee variabelen kennen, kunnen we aan de hand van deze bruto-waarnemingen niets zeggen over de relatie tussen beide kenmerken. Daarom berekenen we vanuit deze bruto-waarnemingen een maat van associatie die de vorm aanneemt van een getal waaruit de sterkte en richting kan worden afgeleid.

Bivariate statistische associatiematen mogen niet zomaar op elke willekeurige set van variabelen worden toegepast. Zoals ook het geval was voor de univariate beschrijvende statistiek, hangt de keuze voor een bivariate beschrijvende associatiemaat af van het *meetniveau* van de twee variabelen in kwestie en van de doelstelling van de onderzoeksfrage. Dit betekent dat de criminoloog nu bij de keuze voor een associatiemaat met een kenmerk meer rekening dient te houden. De keuze voor een bepaalde associatiemaat kan verder bepaald worden door de onderzoeksfrage die de criminoloog zich stelt. Een criminoloog kan twee soorten onderzoeksfragen stellen die kunnen beantwoord worden aan de hand van de bivariate beschrijvende statistiek: *verklarende onderzoeksfragen* of *vragen naar samenhang*, zonder uitspraken te willen doen over de causale relatie zelf.<sup>1</sup> Verklarende onderzoeksfragen hebben doorgaans een causale ondertoon: de onderzoeker vermoedt dat gebeurtenissen van het type A een oorzaak zijn van gebeurtenissen van het type B. Typische verklarende vragen vinden we in de etiologische criminologie. In dat geval wordt de verklarende statistiek

---

<sup>1</sup> Het is zeer belangrijk om niet met de deur in huis te vallen en de statistische associatiematen te presenteren die de criminoloog kan gebruiken wanneer deze onderzoek doet. Met de bivariate beschrijvende statistiek kan de criminoloog geen hypothesen toetsen. Dit wil zeggen: de criminoloog kan geen extrapolatie maken van de onderzoeksbevindingen naar het universum. Dit is het domein van de inferentiële statistiek, en dat deel volgt later in deze syllabus.

gebruikt en wordt de predictie een instrument om een statistische effect-relatie vast te stellen. De effect-relatie kan echter geen waterdicht bewijs vormen voor de causale ondertoon.

## **5. Bivariate frequentieverdelingen voor lage en hoge meetniveaus**

In de bivariate statistiek zijn we geïnteresseerd in bivariate frequentieverdelingen. Een bivariate frequentieverdeling is niet meer dan de frequentieverdeling van twee kenmerken. Aan de hand van een bivariate frequentieverdeling kunnen we nagaan of de frequentieverdeling van één variabele samenloopt met de frequentieverdeling van een ander kenmerk. We geven hier een eenvoudig voorbeeld: is het zo dat we kunnen zeggen dat er een associatie bestaat tussen het gebruik van harddrugs en het betrokken zijn bij georganiseerde diefstal? Er zijn meerdere verklaringen te geven voor de associatie tussen druggebruik en criminaliteit. Voor sommige individuen is het gebruik aanwezig voor de betrokkenheid bij criminaliteit en kan het gebruik een causaal mechanisme zijn. Gebruik vereist immers middelen, en daartoe kan criminaliteit een effectieve strategie zijn. Anderzijds kan een diepe betrokkenheid bij criminaliteit voor de nodige stress zorgen, en kan die stress een factor zijn die maakt dat de regelovertreder occasioneel verdovende middelen gebruikt om de stress te verlichten. Een derde verklaring is nog anders: zowel het druggebruik als de betrokkenheid bij criminaliteit hebben een gemeenschappelijke oorzaak: de zoektocht naar senstatie (thrill-seeking behaviour) en kicks. Hoe men deze associatie wil verklaren, de drie mogelijke scenario's veronderstellen allemaal dat er op zijn minst een associatie bestaat tussen de twee kenmerken.

Bivariate frequentieverdelingen worden voorgesteld aan de hand van een *contingentietabel* (*kruistabel*). Voor metrische variabelen kunnen we ze ook voorstellen aan de hand van een *puntenwolk*, of ook wel *scatterplot* genoemd.

### ***De situatie voor contingentietabellen of kruistabellen***

Voor nominale en ordinale kenmerken worden **contingentietabellen** of **kruistabellen** opgesteld. Een contingentietabel is een tabel waarin de categorieën van twee variabelen tegenover elkaar worden uitgezet en waarin de waargenomen frequentie van elke combinatie van categorieën vermeld staat. Een  $2 \times 2$  tabel is een tabel met twee rijen en twee kolommen. Anders gezegd: er zijn twee categorieën bij elk van beide variabelen. Een  $2 \times 3$  tabel is een tabel met twee rijen en drie kolommen. We zetten dus eerst de rijen en dan de kolommen in de notatie. We illustreren de gebruikelijke terminologie aan de hand van een  $2 \times 2$  tabel voor de variabelen X en Y.

		Variabele X		Totaal
		Cat 1	Cat 2	
Variabele Y	Cat 1	A	B	A+B
	Cat 2	C	D	C+D
Totaal		A+C	B+D	A+B+C+D

A, B, C en D zijn **celfrequenties**: ze geven aan hoeveel keer een bepaalde combinatie van categorieën van variabelen voorkomen.

A+B en C+D noemen we de **rijtotalen** van de contingentietabel.

A+C en B+D noemen we de **kolomtotalen** van de contingentietabel.

A+B+C+D, ook wel de som van alle celfrequenties genoemd, is de **steekproefomvang**.

Een 2\*2 kruistabel: de relatie tussen zelfgerapporteerde winkeldiefstal en cannabisgebruik

			ooit joint gerookt		Totaal
			nooit	ooit	
Winkel-diefstal	nooit	Count	2388	104	2492
		% binnen winkeldiefstal	95.8%	4.2%	100.0%
		% binnen joint gerookt	86.0%	42.6%	82.5%
	ooit	Count	390	140	530
		% binnen winkeldiefstal	73.6%	26.4%	100.0%
		% binnen joint gerookt	14.0%	57.4%	17.5%
Totaal		Count	2778	244	3022
		% binnen winkeldiefstal	91.9%	8.1%	100.0%
		% binnen joint gerookt	100.0%	100.0%	100.0%

We zien in deze kruistabel dat elke cel (elke combinatie tussen de twee categorische variabelen) onderzoekseenheden bevat. We hebben naast de exacte geobserveerde eenheden eveneens de procentuele verdeling duidelijk gemaakt. Je ziet dat de celfrequenties (de aantallen) verschillend zijn voor elke combinatie van categorieën. Absolute cijfers zeggen zo weinig. Daarom presenteren we ook de percentages. Je kan die percentages in twee richtingen berekenen: het aantal jongeren die ooit cannabis hebben gerookt voor jongeren die winkeldiefstal hebben gerapporteerd en voor jongeren die geen winkeldiefstal hebben gerapporteerd. Omgekeerd kan je de percentages winkeldiefstal aflezen voor jongeren die cannabis hebben gerookt en voor jongeren die geen cannabis hebben gerookt. Je kan je nu

afvragen welke de juiste leesrichting is. Dat is een zeer terechte vraag, die we verderop behandelen. Hier volstaat het om vertrouwd te raken met de begrippen celfrequenties, rijtotalen en kolomtotalen.

- De celfrequenties zijn 2388, 104, 390 en 140.
- De rijtotalen zijn 2492 en 530.
- De kolomtotalen zijn 2778 en 244.
- De steekproefomvang is 3022

**Een 3\*2 kruistabel: de relatie tussen dronkenschap en slagen en verwondingen**

			ooit dronken geweest		Total
			nooit	ooit	
Slagen en verwon- dingen	Nooit	Count	1790	358	2148
		%	74.3%	58.2%	71.1%
	Enkele keren	Count	288	108	396
		%	12.0%	17.6%	13.1%
	Drie keer of meer	Count	330	149	479
		%	13.7%	24.2%	15.8%
Total		Count	2408	615	3023
		%	100.0%	100.0%	100.0%

De 2\*2 kruistabel kan ook verder uitgebreid worden naar een r\*k tabel. In het voorbeeld hierboven hebben we een tabel gepresenteerd waar drie rijen te zien zijn en twee kolommen.

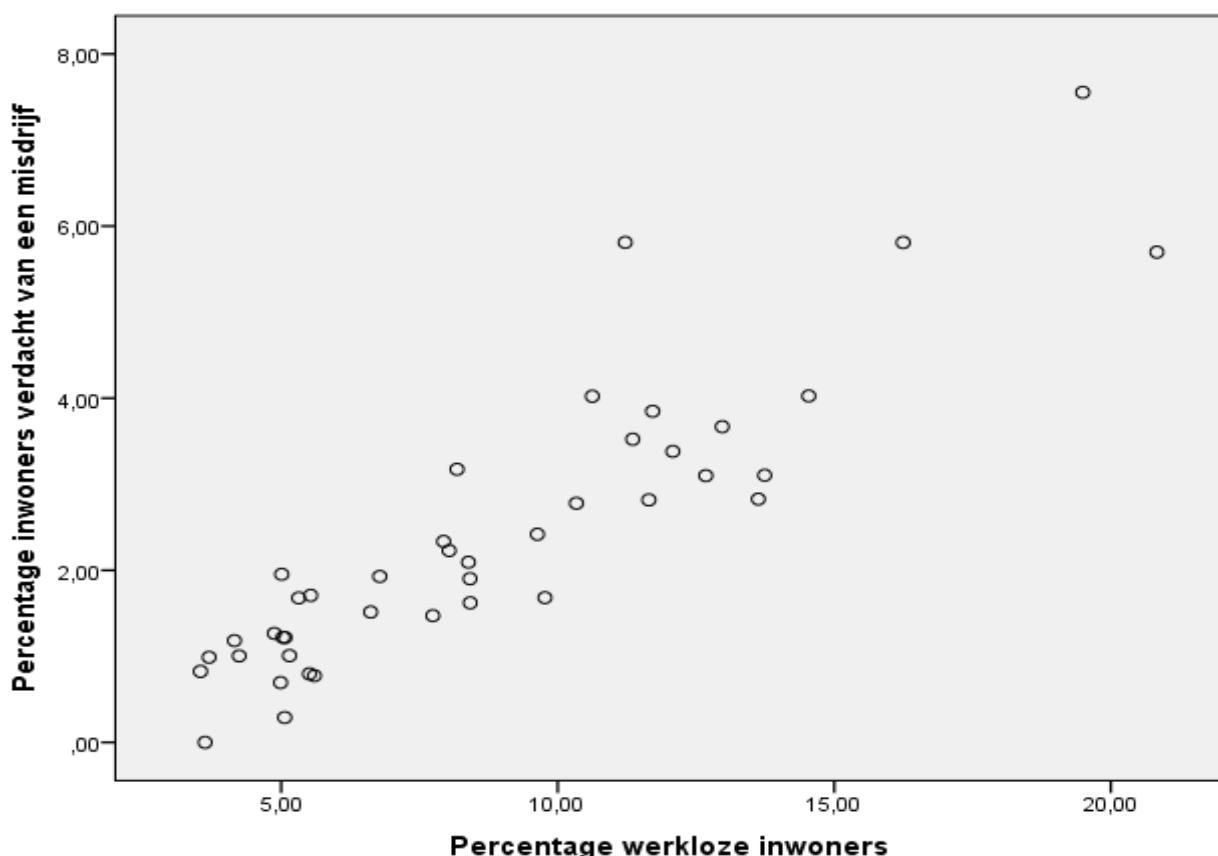
- De celfrequenties zijn nu 1790, 358, 288, 108, 330, 149.
- De rijtotalen zijn nu 2148, 396, 479
- De kolomtotalen zijn hier 2408 en 615
- De totale steekproef bestaat uit 3023 respondenten

### ***De situatie voor metrische variabelen***

Voor metrische variabelen ligt het iets moeilijker om te werken met een contingentietabel. Het zou wel bijzonder onleesbaar zijn om twee metrische variabelen in kruistabel te willen gieten. Het gaat hier immers niet om een beperkt aantal categorieën, maar om variabelen die heel fijn zijn gemeten. Daarom is het allerminst wenselijk om de ruwe data te willen presenteren aan de hand van een kruistabel. Betekent dit dan dat een kruistabel maken niet

wenselijk is? Neen. Het betekent dat we de ruwe gegevens zullen moeten hercoderen willen we de relatie tussen twee metrische kenmerken in een kruistabel gieten. Bij metrische variabelen hebben de onderzoekseenheden heel vaak zeer veel verschillende waarden. Dit komt natuurlijk omdat de variabelen op hoog meetniveau zijn gemeten: bij metrische variabelen is het meetniveau veel fijner en heeft het geen zin om een contingentietabel te maken waar elke waarde op de twee variabelen tegenover elkaar wordt uitgezet. Daarom worden de waarden tegenover elkaar geplaatst op een **scatterplot of puntenwolk**.

Een puntenwolk of scatterplot is een grafische voorstelling van de frequentieverdeling van twee variabelen die gemeten zijn op metrisch niveau. In een puntenwolk worden twee variabelen tegen elkaar uitgezet door de waarden van de variabele X op de **abscis** te plaatsen, en de waarden van de variabele Y op de **ordinaat** te zetten. Conventioneel zetten we altijd de Y-variabele op de ordinaat en de X-variabele onderaan op de abscis. Elke eenheid uit de steekproef kan nu in het **tweedimensionaal** (“er zijn twee assen”) **orthogonaal** (“loodrecht op elkaar staande assen”) vlak worden gesitueerd. Dit gebeurt aan de hand van de waarde van elke eenheid op de variabele X en Y. Deze waarden ( $x_j, y_j$ ) worden de **coördinaten** genoemd.



Dit voorbeeld toont aan hoe we de relatie tussen twee metrische kenmerken presenteren. De puntenwolk is handig, omdat deze ons in een oogopslag al een hint geeft over de relatie tussen de twee kenmerken. Het voorbeeld dat hier gegeven wordt, betreft de relatie tussen twee buurtkenmerken: het percentage werklozen aan de ene kant en het percentage jongeren dat verdacht wordt van een misdrijf aan de andere kant. De gegevens zijn afkomstig van een onderzoek dat in alle Antwerpse buurten werd gevoerd.

## 6. Verantwoord kiezen tussen een reeks van associatiematen

Associatiematen zijn voor de kwantitatieve criminoloog-onderzoeker “*the tools of the trade*” zoals scalpels en andere operatiegereedschappen tot de gereedschapskist van de chirurg behoren. Het komt er steeds op aan de juiste associatiemaat te kiezen. De “juiste” maat is een maat die past bij de **onderzoeksvraag van de criminoloog**. De juiste associatiemaat is een maat die de onderzoeksvraag beantwoordt. Bij de keuze voor een best passende associatiemaat houdt men rekening met **twee regels**:

- *Ten eerste* is er het meetniveau van de afhankelijke en/of de onafhankelijke variabele. **Als beide verschillend zijn, kiest men doorgaans voor het meetniveau van de variabele die op het laagste niveau is gemeten.** Bijvoorbeeld: willen we de relatie nagaan tussen een nominaal en ordinaal kenmerk, dan moeten we een nominale associatiemaat gebruiken.
- *Ten tweede* is er de **verwachte theoretische richting van de associatie**: symmetrische of asymmetrische analysetechnieken worden dan gekozen in functie van de verwachte theoretische associatie. We spreken van een asymmetrische analyse wanneer we uitspraken willen doen in de trant van “X leidt tot Y” of “als X dan Y”. We veronderstellen als het ware een eenrichtingsverkeer; X leidt tot Y maar niet omgekeerd. Om het met een sprekend voorbeeld te stellen: geslacht kan nooit de afhankelijke variabele zijn in criminologisch onderzoek.

## 7. Leerdoelen

In dit hoofdstuk werden een aantal cruciale begrippen behandeld die echt van belang zijn wil men de bivariate verklarende statistiek en bij uitbreiding de multivariate beschrijvende statistiek begrijpen. Deze concepten maken niet enkel het hart uit van de statistiek, maar ook van het kwantitatieve verkennende en theorie-toetsende criminologische onderzoek. Het is dus van belang de betekenis van deze begrippen te kennen aangezien deze begrippen niet

enkel figureren in de statistiek. We verwachten dat de studenten de diverse invullingen die aan het causaliteitsbegrip werden gegeven kennen en vooral ook de beperkingen inzien van elke conceptualisering van causaliteit. De verschillende statistische termen die werden uitgelegd dienen gekend te zijn. Bedenk hierbij best een criminologisch voorbeeld.

## **Hoofdstuk 7**

### **Bivariate associatiematen voor nominale en ordinale variabelen**

#### **1. Inleiding**

In dit hoofdstuk bespreken we de belangrijkste parameters voor de beschrijving van een verband tussen kenmerken van onderzoekseenheden op nominaal en ordinaal niveau. Deze technieken zijn belangrijk, want veel eigenschappen waarin criminologen geïnteresseerd zijn, zijn van het nominale en ordinale niveau.

#### **2. Het percentageverschil als associatiemaat op nominaal niveau**

In een kruistabel of contingentietabel staan de variabelen met hun frequenties paarsgewijs horizontaal en verticaal geplaatst, elk met een aantal (nominale) kenmerken, verdeeld over verschillende categorieën. Kruistabellen zijn tweedimensionale tabellen, verdeeld over kolommen en rijen. We maken een onderscheid tussen  $2 \times 2$  tabellen en  $r \times k$  tabellen, waarbij  $r$  en  $k$  staan voor het aantal rijen en kolommen.  $R \times k$  tabellen zijn een extensie van de eenvoudige  $2 \times 2$  tabel. Bij kruistabellen is het zo dat in elke cel het geobserveerde aantal staat bij een bepaalde combinatie van kenmerken. Kruistabellen worden gebruikt wanneer we de relatie tussen twee nominale kenmerken willen bestuderen. De verschillen in de afhankelijke variabele worden vergeleken over de verschillende categorieën van de onafhankelijke variabele. We gebruiken hiervoor vaak het *percentageverschil* en de *odds ratio*.

Stel dat we onderzoek doen naar de betrokkenheid bij een problematische jeugdgroep en geslacht. We willen weten of jongens in sterkere mate betrokken zijn bij een problematische jeugdgroep dan meisjes. In een kruistabel kunnen we deze informatie eenvoudig weergeven. De samenhang tussen twee nominale variabelen kunnen we omschrijven als de manier waarop een kenmerk verdeeld is binnen de categorieën van een ander kenmerk. In het voorbeeld kijken we dus naar de manier waarop betrokkenheid bij geweld in groepsverband verdeeld is naar geslacht in een eenvoudige  $2 \times 2$  tabel.

**Betrokkenheid bij een problematische jeugdgroep en geslacht**

	Geslacht		Totaal
	Meisjes	Jongens	
Betrokkenheid bij problematische jeugdgroep	Niet betrokken	1208	1120
	Betrokken	45	101
Totaal		1253	1221
			2328
			146
			2474

1208, 1120, 45 en 101 noemen we de **celfrequenties**: ze geven aan hoeveel keer een bepaalde combinatie van categorieën voorkomt. (1208+1120) of 2328 en (45+101) of 146 noemen we de **rijtotalen**. (1208+45) of 1253 en (1120+101) of 1221 noemen we de **kolomtotalen**. Rijtotalen en kolomtotalen worden ook wel de **marginalen** genoemd. De marginalen geven de univariate frequentieverdeling van de respectievelijke variabelen weer. De som van alle celfrequenties is het totaal aantal waarnemingen, de *populatieomvang* of de *steekproefomvang*. Aangezien het interpreteren van absolute aantallen in de categorieën vaak moeilijk is (doordat de groepen die we vergelijken niet hetzelfde totaal aantal hebben), moeten de absolute aantallen omgerekend worden naar proporties of percentages (herinner je: percentages zijn proporties vermenigvuldigd met honderd). Zeker indien we de associatie tussen twee variabelen wensen te vergelijken, is het werken met proporties of percentages noodzakelijk.

Een belangrijk probleem hierbij is de *richting waarin we percenteren*. We kunnen binnen de rijen, binnen de kolommen of ten opzichte van het totaal aantal waarnemingen percenteren. De richting van de associatie wordt bepaald door theoretische verwachtingen die men heeft. Statistisch dienen we bij het maken van een bivariate kruistabel het onderscheid te maken tussen een *onafhankelijke* en een *afhankelijke* variabele.

- Een afhankelijke variabele is een kenmerk dat beïnvloed wordt door één of meerdere andere kenmerken.
- Een onafhankelijke variabele is een kenmerk dat invloed uitoefent op een ander kenmerk.

In het voorbeeld beschouwen we geslacht als een onafhankelijke variabele en het betrokken zijn bij een problematische jeugdgroep als een afhankelijke variabele. Het moge nogmaals duidelijk zijn dat geslacht nooit een afhankelijke variabele kan zijn. Of men jongen of meisje is, wordt niet bepaald door het lidmaatschap van een problematische jeugdgroep. Dit is een

inhoudsloze uitspraak. Werken met kruistabellen impliceert, zoals elke bivariate (en multivariate) analyse, een gedegen criminologische argumentatie waarom bepaalde kenmerken als afhankelijk dan wel als onafhankelijk worden beschouwd. In het onderstaande voorbeeld beschouwen we geslacht als onafhankelijke variabele. We hebben de gewoonte om de afhankelijke variabele in de rijen te plaatsen en de onafhankelijke variabele in de kolommen. Dit is echter niet verplicht. Als de afhankelijke variabele in de rijen wordt gepresenteerd, dan dienen kolompercentages gemaakt te worden en dienen deze percentages te worden vergeleken voor de verschillende categorieën van geslacht. Als de afhankelijke variabele in de kolommen wordt geplaatst, dienen de rijpercentages worden berekend. We gaan in dit handboek echter steeds de afhankelijke variabele in de rijen plaatsen.

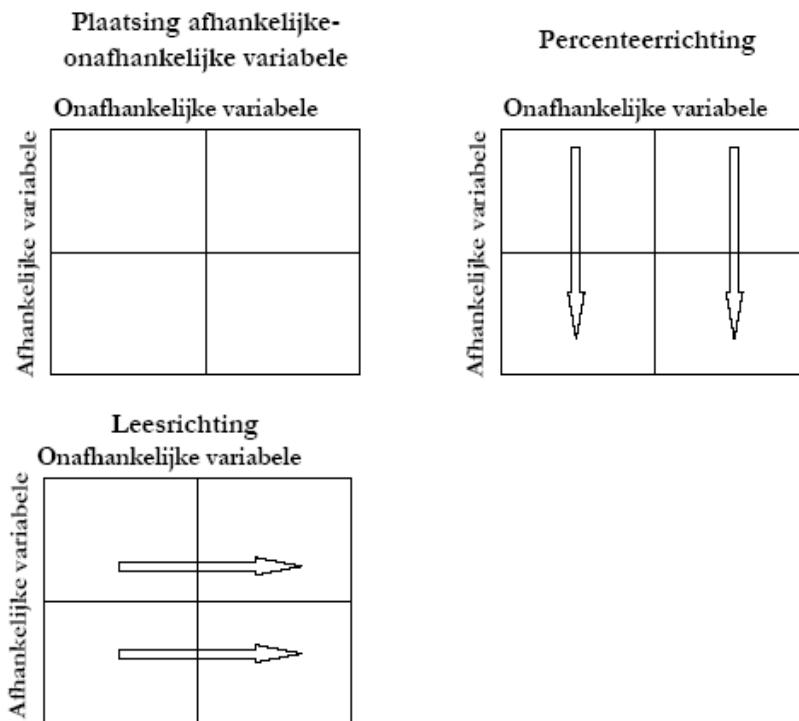
**Betrokkenheid bij een problematische jeugdgroep en geslacht**

		geslacht		Totaal
		meisje	jongen	
Betrokkenheid bij problematische jeugdgroep	Niet betrokken	Absoluut aantal	1208	1120
		Kolompercentage binnen geslacht	96,4%	91,7%
		Absoluut aantal	45	101
	Betrokken	Kolompercentage binnen geslacht	3,6%	8,3%
		Absoluut aantal	1253	1221
		Kolompercentage binnen geslacht	100,0%	100,0%
Totaal			2328	146
			5,9%	2474
			100,0%	100,0%

In het voorbeeld zien we dat slechts 3.6% van de meisjes lid is van een problematische jeugdgroep en 8.3% van de jongens. Dit geeft een verschil van **4.7 percentagepunten**. De logica van een analyse van kruistabellen is helder: bekijk de relatieve frequenties van de variabele “betrokkenheid bij een problematische jeugdgroep” nu eens als waren het twee aparte univariate frequentieverdelingen. Dan zie je dat 3.6% van de meisjes betrokken is bij een problematische jeugdgroep en dan zie je dat 8.3% van de jongens uit de steekproef betrokken is bij een problematische jeugdgroep. Stel nu eens dat er helemaal geen associatie was tussen de twee kenmerken. Wat zouden we dan verwachten? Weet je het niet? Volg dan intuïtief je verstand en redeneer mee. Als er geen verband zou bestaan tussen de beide kenmerken, dan zou het toch logisch zijn dat de frequentieverdelingen gelijk lopen? Als er geen verband bestaat, dan verwachten we dat de situatie bij jongens en bij meisjes niet zou afwijken van de totale frequentieverdeling: het percentage van 5.9% zou dan zowel bij meisjes als bij jongens moeten geobserveerd worden. Maar dat doen we niet. We observeren dat de percentages verschillend zijn, en wel in het nadeel van de jongens. Jongens hebben dus

een grotere kans om betrokken te zijn bij een problematische jeugdgroep. Het verband tussen geslacht en betrokkenheid bij problematische jeugdgroepen, gangs, bendes, georganiseerde misdaad, enz. is bijzonder stabiel. De meeste studies tonen dergelijk verband aan. Hoe we het verband dienen te interpreteren, is nog maar de vraag. Geslacht op zich is geen oorzakelijk mechanisme dat een gebeurtenis kan teweegbrengen, maar geslacht kan wel een markeerde zijn voor een ander ongemeten kenmerk, zoals het testosteron gehalte. Het kan ook zijn dat er geen biologische maar een sociaalpsychologisch mechanisme aan de basis ligt van de geobserveerde samenhang. Mogelijks is het zo dat er minder toezicht is op jongens dan op meisjes en is het zo dat jongens er een meer risicovolle levensstijl op na houden die hen meer met bendes in contact brengt. Hierdoor hebben ze een grotere kans om door bendes gerekruteerd te worden. De cijfers spreken dus niet voor zichzelf. De criminoloog dient de interpretatie te maken en dat is vaak moeilijker dan de eenvoudige berekening van een statistisch verband.

Laten we nu nog eenmaal terugkeren op de richting voor het percenteren omdat studenten daar vaak tegen zondigen. Onderstaande figuur vat de wijze waarop kruistabellen dienen gelezen te worden nog eens samen.



De wijze waarop we percenteren is volledig afhankelijk van de plaats die afhankelijke en onafhankelijke variabelen in een tabel innemen. Het is niet steeds mogelijk om onafhankelijke en afhankelijke variabelen van elkaar te onderscheiden. Dit is het geval waar twee variabelen worden vergeleken en waarbij de ene de verklaring kan vormen voor de andere en omgekeerd. Het valt dus wel eens voor dat er theoretische discussie bestaat over de richting van de relatie tussen twee kenmerken. Nemen we bijvoorbeeld religie en onverdraagzaamheid: het is mogelijk dat het aanhangen van een bepaalde religie leidt tot een hogere mate van onverdraagzaamheid. Anderzijds kan het evengoed zijn dat onverdraagzame mensen een bepaalde soort religie gaan aanhangen. In dat geval dient de criminoloog een keuze te maken en die te verduidelijken. Dit kan door bijvoorbeeld te verwijzen naar wat er in de literatuur gezegd wordt.

Naast de richtlijnen bij het percenteren van de tabellen dienen volgende regels te worden gevuld:

1. Vermeld altijd het *totaal aantal waarnemingen* indien enkel de percentages of proporties in een tabel werden opgenomen.
2. Bereken *nooit een percentage wanneer het aantal gevallen kleiner is dan 30*. Wanneer we de waarnemingen in een bepaalde categorie met één eenheid verminderen dan betekent dit een verlaging met 3%. Terwijl een verschil van 3% bij een groot aantal waarnemingen significant mag genoemd worden, is dit bij minder dan 30 waarnemingen niet meer het geval. De kans op foute interpretaties wordt dan ook heel groot. In dit geval is het beter enkel de absolute waarden weer te geven.

### **3. De odds ratio als associatiemaat op nominaal niveau**

Een andere manier om de relatie tussen twee nominale kenmerken te bestuderen is door de **odds ratio** te berekenen. De odds is een maat om de *verhouding tussen het voorkomen van een gebeurtenis en het niet voorkomen van een gebeurtenis* uit te drukken. (vb. Het aantal 12 jarigen in de steekproef dat overgaat tot het plegen van een delict gedeeld door het aantal 12 jarigen in die steekproef dat niet overgaat tot het plegen van een delict). De odds ratio is de verhouding tussen twee odds en wordt ook wel de *kruisproduct ratio* genoemd. De odds ratio is een **asymmetrische associatiemaat** die heel gemakkelijk te berekenen is. Verder is het zo dat de odds ratio een aantal heel interessante eigenschappen heeft. Odds ratio's zijn namelijk niet zo gevoelig voor de marginale verdelingen. We merkten hierboven op dat het percenteren

niet zonder gevaren is in kleine steekproeven. De odds ratio kan een alternatief zijn. De odds ratio is tevens de basis voor de logistische regressie-analyse die verder nog aan bod komt. De interpretatie van de odds ratio is relatief eenvoudig. **De odds ratio neemt de waarde aan van 1 bij afwezigheid van een verband en wijkt af van 1 naarmate het verband sterker wordt. De afwijking gebeurt naar 0 of naar + oneindig.**<sup>6</sup>

Hernemen we de hierboven gepresenteerde tabel over de relatie tussen geslacht en betrokkenheid bij een problematische jeugdgroep, dan kunnen we de verhoudingen tussen het betrokken zijn en niet betrokken zijn eerst afzonderlijk berekenen voor jongens en meisjes. Deze afzonderlijke verhoudingen noemen we **odds**.

*Uitgewerkt rekenvoorbeeld:*

		geslacht		Totaal
		meisje	jongen	
Betrokkenheid bij problematische jeugdgroep	Geen betrokkenheid	1208	1120	2328
	Betrokkenheid	45	101	146
Totaal		1253	1221	2474

De verhouding tussen het niet-betrokken zijn bij een problematische jeugdgroep en het betrokken zijn bij een problematische jeugdgroep (of de odds voor het niet-betrokken zijn bij een problematische jeugdgroep) ziet er voor meisjes als volgt uit: (1208/45) of 26,84. *Dit betekent dat meisjes 26,84 keer meer kans hebben om niet betrokken te zijn bij een problematische jeugdgroep dan wel betrokken te zijn bij een problematische jeugdgroep.* Onder jongens ziet deze verhouding er als volgt uit: (1120/101) of 11,08.

Wanneer we nu de verhouding nemen van deze twee verhoudingen, kennen we de relatieve verhouding naar geslacht. De odds ratio of de verhouding tussen twee odds naar geslacht ziet

---

<sup>6</sup> De presentatie van odds kan verwarringen opleveren omdat de waarde een verband en omdat een odds-ratio niet negatief kan zijn. Dit kan worden opgelost door met je rekenmachine de odds-ratio om te zetten in een logaritmische schaal (natuurlijk logaritme). Het natuurlijk logaritme van 1 is nul, en nul wordt dan het nulpunt en betekent geen verband. Odds-ratio's die lager zijn dan 1, hebben een negatieve log-odds waarde en odds-ratio's die hoger zijn dan 1 hebben een positieve waarde. De multivariate analysetechnieken voor categorische data-analyse, zoals de logistische regressieanalyse, is gebaseerd op de odds-ratio's en log-odds. Dit is niet belangrijk in een verkennend handboek, waar enkel de basis gekend dient te zijn, maar in latere jaren wordt deze materie zeker behandeld.

er als volgt uit:  $26,84 / 11,08$  of  $2,4$ . Dit betekent dat meisjes 2,4 keer meer kans hebben om niet betrokken te zijn bij een problematische jeugdgroep dan jongens. Er is dus een associatie.

Bij wijze van voorbeeld tonen we in onderstaande tabel een reeks van risico- en beschermingsfactoren en hun relatie met regelovertredend gedrag. De resultaten zijn afkomstig uit de zelfrapportagestudie van eerstegraadsleerlingen te Antwerpen, waarbij 2486 jongeren werden bevraagd over hun regelovertredend gedrag. Odds-ratio's worden vaak gepresenteerd in onderzoek naar **risicofactoren** van regelovertredend gedrag. De odds-ratio is dus een belangrijke maat in criminologisch onderzoek.

De tabel is een goede manier om enerzijds verbanden te leren interpreteren en aflezen en anderzijds na te denken over de betekenis van de verbanden. Alle variabelen zijn gebaseerd op meerdere indicatoren (vragen) en de antwoorden werden daarna herleid tot twee categorieën: hoog of laag. Neem het voorbeeld van de werkloosheid van de moeder. Je ziet dat er een positief verband bestaat tussen de werkloosheid van de moeder en ernstige delinquentie (OR van 1.56). Jongeren met een werkloze moeder hebben 1.56 meer kans om ernstig delinquent gedrag te vertonen dan jongeren die geen werkloze moeder hebben. Het verband is er wel, maar is lang niet zo sterk als andere verbanden die in de tabel af te lezen zijn. Neem nu lage zelfcontrole: jongeren met een lage zelfcontrole (dus jongeren die er niet goed in slagen zichzelf te beheersen) hebben 4.15 keer meer kans op het vertonen van ernstige delinquentie dan jongeren met een hoge zelfcontrole. Bekijk de tabel eens en tracht voor jezelf en tracht alle verbanden te lezen. Welke vind je zelf opmerkelijk? Welke verbanden had je niet verwacht?

**Odds-Ratio's voor ernstige delinquentie, gewelddadige delinquentie, veelplegers, niet-plegers en betrokkenheid bij een gewelddadige jeugdgroep (VYG)**

Risicofactoren of beschermingsfactoren	Ernstige delinquentie	Gewelddadige delinquentie	Veelplegers	Betrokkenheid bij een problematische jeugdgroep
<b>Jongen</b>	<b>2.84</b>	<b>2.78</b>	<b>2.83</b>	<b>2.12</b>
<b>Beide ouders Belgisch</b>	<b>0.48</b>	<b>0.48</b>	<b>0.54</b>	<b>0.44</b>
<b>Blijven zitten</b>	<b>1.97</b>	<b>1.51</b>	<b>2.02</b>	<b>2.00</b>
<b>Eenoudergezin</b>	1.24	1.17	<b>1.53</b>	1.29
<b>Werkloosheid vader</b>	<b>2.84</b>	<b>1.38</b>	1.21	1.11
<b>Werkloosheid moeder</b>	<b>1.56</b>	<b>1.28</b>	<b>1.48</b>	1.19
<b>Armoede</b>	<b>1.41</b>	1.18	1.33	<b>1.65</b>
<b>Residentiële stabiliteit buurt</b>	<b>0.61</b>	<b>0.71</b>	<b>0.65</b>	<b>0.56</b>
<b>Ooit opgepakt door politie</b>	<b>7.65</b>	<b>5.77</b>	<b>10.92</b>	<b>5.48</b>
<b>Lid gewelddadige jeugdgroep</b>	<b>7.51</b>	<b>6.43</b>	<b>13.40</b>	--
<b>Ouderlijke gehechtheid</b>	<b>0.49</b>	<b>0.48</b>	<b>0.37</b>	<b>0.34</b>
<b>Ouderlijk toezicht</b>	<b>0.33</b>	<b>0.39</b>	<b>0.18</b>	<b>0.15</b>
<b>Integratie in de klas</b>	0.85	0.95	1.04	0.76
<b>Studiebetrokkenheid</b>	<b>0.37</b>	<b>0.43</b>	<b>0.22</b>	<b>0.36</b>
<b>Lage moraliteit</b>	<b>4.66</b>	<b>4.91</b>	<b>10.34</b>	<b>8.82</b>
<b>Externe locus of control</b>	<b>2.46</b>	<b>2.39</b>	<b>3.10</b>	<b>3.88</b>
<b>Impulsiviteit</b>	<b>3.43</b>	<b>3.70</b>	<b>6.98</b>	<b>5.37</b>
<b>Woedebeheersing</b>	<b>3.27</b>	<b>3.81</b>	<b>4.25</b>	<b>3.37</b>
<b>Relatieve depravatie</b>	1.00	<b>1.21</b>	0.89	0.95
<b>Criminale geneigdheid</b>	<b>5.80</b>	<b>7.51</b>	<b>16.71</b>	<b>14.32</b>
<b>Lage zelfcontrole</b>	<b>4.15</b>	<b>4.73</b>	<b>9.13</b>	<b>5.58</b>
<b>Criminale leeftijdsgenoten</b>	<b>3.98</b>	<b>4.85</b>	<b>7.89</b>	<b>7.88</b>
<b>Ongestructureerde vrijetijd</b>	<b>2.65</b>	<b>2.85</b>	<b>4.40</b>	<b>3.65</b>
<b>Risicotvolle leefstijl</b>	<b>5.51</b>	<b>5.46</b>	<b>12.14</b>	<b>11.21</b>

Bron: Pauwels (2007) de vetgedrukte resultaten zijn statistisch significant ( $p < 0.05$ )

#### 4. Chi-kwadraat ( $\chi^2$ ) als associatiemaat op nominaal niveau

Een **chi-kwadraattoets** wordt in de statistiek gebruikt om te zien of waargenomen cel frequenties systematisch afwijken van verwachte cel frequenties indien geen associatie zou bestaan tussen twee kenmerken. Een chi-kwadraattoets wordt veel gebruikt om kruistabellen te analyseren. De chi-kwadraat kan beschouwd worden als een maat voor de sterkte van een

relatie tussen twee variabelen gemeten op **nominaal** meetniveau, of tussen een nominale en een ordinale variabele. De waarde van de chi-kwadraat neemt toe naarmate de associatie tussen de variabelen sterker is. Het is echter geen eenduidig te interpreteren maat. Het is m.a.w. moeilijk deze associatiemaat te interpreteren. *De chi-kwadraat varieert van 0 bij afwezigheid van een verband tot zeer hoge waarden.* Er is **geen absolute begrenzing** aan de waarden die chi-kwadraat kan aannemen, vandaar dat het moeilijk is de sterkte van het verband te interpreteren. Daarom zijn er door statistici op chi-kwadraat gebaseerde associatiematen bedacht, zoals ***Phi***, de **contingentiecoëfficient C** en **Cramer's V**, die eenvoudiger te interpreteren zijn.

Chi-kwadraat is gebaseerd op een vergelijking van twee kruistabellen. De eerste kruistabel bestaat uit de **geobserveerde frequenties (de werkelijke waarden)**, deze wordt vergeleken met een hypothetische tabel (of het theoretische model) waarin de **verwachte frequenties bij statistische onafhankelijkheid** worden berekend. Dit vereist een beetje meer uitleg. Wat bedoelen we met deze beide begrippen? Het is cruciaal dat je ze kent en goed begrijpt, anders slaag je er niet in de chi-kwadraat waarde uit te rekenen. We gaan hier uit van een redenering die een beetje te vergelijken valt met de situatie van het percentageverschil: we hebben daar vastgesteld dat er een verband is tussen twee variabelen als de frequentieverdeling voor de beide categorieën van de onafhankelijke variabelen niet dezelfde was: er was een verschil tussen jongens en meisjes in termen van hun betrokkenheid bij een problematische jeugdgroep. Als er geen verband zou zijn tussen de beide kenmerken, dan verwachten we dat de totale frequenties niet verschillend zijn bij jongens en meisjes. Maar dat waren ze in het voorbeeld dus duidelijk wel. Welnu, bij de berekening van chi-kwadraat passen we opnieuw een analoge redenering toe. We vertrekken vanuit de geobserveerde celfrequenties en gaan deze situatie vergelijken met de verwachte celfrequenties die we zouden vinden in de hypothetische situatie dat er geen verband bestaat tussen de beide kenmerken. Het komt er dus op aan die “verwachte celfrequenties in de situatie dat er geen verband bestaat tussen beide kenmerken” zelf te gaan berekenen.

Als je deze waarde hebt gevonden, kan je die waarde in rekening brengen en het verschil berekenen tussen de werkelijke waarde of de geobserveerde waarde in een cel en de verwachte waarde in de situatie dat er geen verband is. De wetenschappelijke benaming hiervoor is de verwachte situatie bij statistische onafhankelijkheid. De verwachte frequenties bij statistische onafhankelijkheid worden berekend op basis van de marginalen van de

geobserveerde waarden en de frequenties in de tabel worden zodanig verspreid over de verschillende waarden dat beide variabelen geen enkele associatie hebben. De formule voor de berekening van chi-kwadraat ziet er als volgt uit:

$$\chi^2 = \sum \frac{(f(o)_{ij} - f(e)_{ij})^2}{f(e)_{ij}} \quad \chi^2 = \sum \frac{(geobserveerd - verwacht)^2}{verwacht}$$

Chi-kwadraat is evenwel erg gevoelig voor het aantal meeteenheden in onze tabel. Als het aantal respondenten in een steekproef verdubbelt, dan verdubbelt ook de chi-kwadraat waarde in vergelijking met de oorspronkelijke chi-kwadraat. Daarom wordt aangeraden een sterktemaat te berekenen die de chi-kwadraat ongevoelig maakt voor de grootte van de steekproef.

Opgelet! Om **chi-kwadraat** te kunnen gebruiken, dienen een aantal **voorwaarden** vervuld te zijn:

- **Ten eerste**, de data in de contingentietabel dienen ruwe frequenties te zijn, geen scores noch percentages.
- **Ten tweede**, de onderzochte variabelen dienen categorisch te zijn en de meetwaarden dienen elkaar uit te sluiten. Elke observatie (meeteenheid) mag slechts in één cel thuishoren.
- **Ten derde**, de  $\chi^2$  waarde mag maar geïnterpreteerd worden indien aan een aantal voorwaarden is voldaan. Deze houden in dat maximaal 20% van de cellen een verwachte frequentie bevatten van  $< 5$ , en geen enkele een verwachte frequentie van 0. Indien dit wel het geval is, dienen cellen samenge trokken te worden.<sup>7</sup>

---

<sup>7</sup> Statistische verwerkingspakketten zoals SPSS geven ons altijd een waarschuwing of aan deze voorwaarde voldaan is: 0 cells (0%) have expected count less than 5. The minimum expected count is 8,50.

### Betrokkenheid bij problematische jeugdgroep: geobserveerde versus verwachte celfrequenties

			Geslacht		Totaal
			Meisjes	Jongens	
Betrokkenheid bij problematische jeugdgroep	Niet betrokken	geobserveerd	1185	1090	2275
		verwacht	1151,5	1123,0	2275,0
		% binnen geslacht	94,6%	89,2%	91,9%
Betrokken	Betrokken	geobserveerd	68	132	200
		verwacht	101,5	99,0	200,0
		% binnen geslacht	5,4%	10,8%	8,1%
Totaal		geobserveerd	1253	1222	2475
		verwacht	1253,0	1222,0	2475,0
		% binnen geslacht	100,0%	100,0%	100,0%

### Een uitgewerkt rekenvoorbeeld

Indien er geen associatie zou zijn tussen het al dan niet betrokken zijn bij een problematische jeugdgroep en geslacht, zouden een even grote proportie jongens als meisjes niet betrokken zijn bij een problematische jeugdgroep en een even grote proportie jongens als meisjes wel betrokken zijn bij een problematische jeugdgroep. Hiervoor kijken we naar de laatste kolom met de totalen.

- We zien dat er 91,9% van de respondenten (ongeacht het geslacht) niet betrokken is bij een problematische jeugdgroep.
- Bij afwezigheid van associatie zouden we dus kunnen verwachten dat zowel 91,9% van de meisjes als 91,9% van de jongens niet betrokken is bij een problematische jeugdgroep.
- De verwachte waarde (bij statistische onafhankelijkheid) voor het **niet betrokken** zijn bij een problematische jeugdgroep **bij meisjes** berekenen we als volgt:  
**91,9%\*1253= 1151,5**
- De verwachte waarde (bij statistische onafhankelijkheid) voor het **niet betrokken** zijn bij een problematische jeugdgroep **bij jongens** berekenen we als volgt:  
**91,9%\*1222= 1123,0**
- De verwachte waarde (bij statistische onafhankelijkheid) voor het **betrokken** zijn bij een problematische jeugdgroep **bij meisjes** berekenen we als volgt:  
**8,1%\*1253= 101,5**

- De verwachte waarde (bij statistische onafhankelijkheid) voor het **betrokken** zijn bij een problematische jeugdgroep **bij jongens** berekenen we als volgt:

$$8.1\% * 1222 = 99.0$$

**Chi kwadraat** wordt dan zo berekend, we nemen de formule er nog eens bij:

$$\chi^2 = \sum \frac{(geobserveerd - verwacht)^2}{verwacht}$$

$$\begin{aligned} \text{Chi-kwadraat} &= ((1185-1151.5)^2 / 1151.5) + ((1090-1123.0)^2 / 1123.0) + ((68-101.5)^2 / 101.5) \\ &+ ((132-99)^2 / 99) = 0.97 + 0.97 + 11.06 + 11 = 24,00 \end{aligned}$$

De chi-kwadraat waarde bedraagt voor deze kruistabel 24. **Dit moet je als volgt interpreteren:** er is wel degelijk een verschil tussen de geobserveerde celfrequenties en de verwachte celfrequenties indien er geen associatie zou bestaan tussen beide kenmerken. **Je voelt intuïtief aan dat je hier niet echt wijzer van wordt. Het is immers moeilijk een verband te interpreteren zonder richtlijnen. Richtlijnen helpen omdat ze bijvoorbeeld de bovengrens en ondergrens aangeven van een verband. Chi-kwadraat heeft geen vaste bovengrens.** Indien onze steekproef dubbel zo groot zou zijn, zou de waarde van chi-kwadraat bij eenzelfde associatie verdubbelen. Hoe moeten we nu precies vaststellen hoe sterk de associatie is? Als er wel degelijk een verband bestaat, dan moet dat verband onafhankelijk van de steekproefgrootte kunnen worden bepaald. Het mag er niet toe doen of de steekproef nu uit vijfhonderd respondenten bestaat of uit duizend respondenten. Statistici kenden dit probleem en hebben daar een handige oplossing voor bedacht. Ze vonden een manier om de associatiemaat chi-kwadraat te gaan normeren. De mogelijkheden om te normeren zijn niet oneindig, maar een handige manier om daar toch in te slagen is om de chi-kwadraat waarde te gaan relateren aan de grootte van de steekproef. Op die manier is het alvast niet meer zo dat de steekproefgrootte het verband gaat beïnvloeden. Er zijn twee varianten bedacht van associatiematen die allebei op chi-kwadraat gebaseerd zijn: de ene associatiemaat is Phi en de andere associatiemaat is Cramer's V. Deze maten corrigeren voor de problemen die aanwezig zijn bij het gebruik van chi-kwadraat als associatiemaat.

## 5. Phi

**Phi** is een associatiemaat die gebaseerd is op chi-kwadraat en neemt de *waarde nul aan bij geen associatie en de waarde van één bij een perfecte statistische associatie*. Phi wordt gebruikt bij de berekening van de associatie tussen kenmerken in een 2\*2 tabel.

De formule ziet er als volgt uit:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Toegepast op ons voorbeeld:

$$\sqrt{\frac{24}{2475}} = 0.0985 \\ = 0.1$$

In het voorbeeld hierboven bedraagt phi 0.1. Het gaat hier dus om een zwakke associatie.

## 6. Cramer's V

**Cramer's V** is o.i. de meest aangewezen coëfficiënt van die associatiematen die steunen op chi-kwadraat. Cramer's V is belangrijk want deze associatiemaat corrigeert voor de gevoeligheid van chi-kwadraat voor de steekproefgrootte en het aantal rijen en kolommen van variabelen. Cramer's V varieert eveneens van nul tot één en wordt gebruikt bij de berekening van associaties tussen kenmerken in een r\*k tabel. In het voorbeeld hierboven bedraagt Cramers's V 0.1. Het gaat hier dus om een zwakke associatie.

Volledigheidshalve geven we ook de formule weer:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \\ (k = \min(r, k))$$

k = 2 (want 2\*2 tabel)

$$v = \sqrt{\frac{24}{2475(2-1)}} = 0.0985$$

V= 0.1

## 7. Gamma als associatiemaat op ordinaal niveau

**Gamma** wordt nog regelmatig gebruikt om de samenhang tussen variabelen van het ordinale niveau te berekenen. De waarde van gamma verloopt van *min één tot plus één*, waarbij min één wijst op een perfect negatief verband, plus één op een perfect positief verband en nul op de afwezigheid van een verband. Hoe verder van nul, hoe sterker de associatie!

Heeft de associatiemaat een waarde kleiner dan nul, dan betekent het dat hoge waarden op de ene variabele samengaan met lage waarden op de andere variabele. We spreken van een **negatieve relatie**. Heeft de associatiemaat een waarde groter dan nul, dan betekent het dat hoge waarden op de ene variabele samengaan met hoge waarden op de andere variabele. Er is sprake van een **positieve relatie**.

Ordinale variabelen, zoals vragen uit een vragenlijst waarbij de antwoordcategorieën variëren van “helemaal niet akkoord” tot “helemaal akkoord” of waarbij de antwoordcategorieën variëren van “nooit” over “soms”, “vaak” tot “zeer vaak”, zijn eigenlijk strikt genomen van het ordinale niveau: de afstand tussen helemaal niet akkoord en helemaal akkoord is niet dezelfde als de afstand tussen akkoord en helemaal akkoord.<sup>8</sup> Beschouw onderstaande 3\*3 kruistabel. **Gamma** is gebaseerd op de verhouding tussen *consistente paren* en *inconsistente paren*. **Een consistent paar** is ieder paar van elementen waarbij de waarde van het ene element op beide variabelen hoger is dan van het andere element. **Een inconsistent paar** is ieder paar van elementen waarbij de waarde van het ene element op de ene variabele hoger is en op de andere variabele lager is dan van het andere element.

---

<sup>8</sup> Hoewel numerieke waarden gebruikt worden bij de invoer van de gegevens in de gegevensmatrix, vormen deze numerieke waarden geen representatie van een echte metrische waarde en daarom dient eigenlijk een ordinale associatiemaat berekend te worden om het verband te kennen tussen twee ordinale kenmerken.

**Tabel: Consistente en inconsistente paren bij de berekening van gamma**

Delinquent gedrag	Sociaal economische status		
	Laag	Midden	Hoog
Laag	A      B      C		
Midden	D      E      F		
Hoog	G      H      I		

Wanneer we het aantal consistente paren en inconsistente paren in bovenstaande 3\*3 tabel identificeren, bekomen we het volgende:

**CP (Consistente paren) zijn :**  $(A*E) + (A*F) + (A*H) + (A*I) + (B*F) + (B*I) + (D*H) + (D*I) + (E*I)$

**IP (Inconsistentie paren) zijn :**  $(C*E) + (C*D) + (C*H) + (C*G) + (B*D) + (B*G) + (F*H) + (F*G) + (E*G)$

Gamma wordt als volgt berekend: **(CP-IP) / (CP + IP)**. Dit getal situeert zich tussen min één en plus één. Dit is de waarde van gamma. Laten we een voorbeeld geven uit criminologisch onderzoek. De waarde van Gamma moet je niet met de rekenmachine zelf kunnen uitrekenen.

**Kruistabel: regels zijn gemaakt om te breken en spijbelen van vrienden**

		Regels zijn gemaakt om te breken					Totaal
		helemaal oneens	beetje oneens	noch eens, noch oneens	beetje eens	helemaal eens	
Spijbelen van vrienden	Geen vriend	811	261	265	185	94	1616
	Eén vriend	221	154	159	130	83	747
	Twee vrienden	11	10	15	16	15	67
	Drie of meer vrienden	9	5	9	4	11	38
Totaal		1052	430	448	335	203	2468

Deze kruistabel bevat twee variabelen uit een zelfrapportagestudie onder jongeren. We vroegen aan de jongeren in welke mate ze het eens waren met de uitspraak “*regels zijn gemaakt om te breken*”. Deze uitspraak is een indicator voor lage moraliteit. We vroegen ook aan diezelfde jongeren hoeveel van hun beste vrienden spijbelen. Wat verwachtten de onderzoekers nu? Wel, er wordt in de criminologische literatuur nogal eens gewezen op het

verband tussen leeftijdsgenoten en eigen waarden en normen. Je kan dat verband op twee manieren interpreteren: leeftijdsgenoten beïnvloeden de eigen waarden en normen, maar je zou ook het omgekeerde kunnen verwachten: de eigen waarden en normen spelen mee bij de keuze voor de eigen vriendenkring. In beide gevallen verwachten we echter dat de bivariate samenhang positief is: dat wil zeggen: hoge waarden op de ene variabele moeten samengaan met hoge waarden op de andere variabele. Anders gezegd: hoe positiever men staat tegenover regelovertreding, hoe meer spijbelende vrienden men heeft. Het probleem van de kip of het ei lossen we met deze statistische analyse niet op. Wel kunnen we vaststellen of de idee van samenhang klopt. Deze blijkt wel degelijk te kloppen. Gamma bedraagt hier 0.328. Er is dus een matig verband tussen de ordinale variabelen “regels zijn gemaakt om te breken” en het aantal spijbelende vrienden. Gamma kan ook berekend worden voor metrische variabelen die men ordinaal maakt door deze variabelen in categorieën te brengen. Echter, het is mogelijk dat de samenhang gebaseerd op de maat gamma ietwat verschilt van de samenhang die gebaseerd is op de metrische correlatiecoëfficiënt. Dat is mogelijk omdat we informatie verliezen: als we variabelen hercoderen en een zeer gedetailleerde metrische variabele herleiden tot een ordinale variabele met vijf categorieën, gooien we toch wat informatie weg. Als we metrische variabelen in categorieën onderverdelen, moeten we dat op een verstandige manier doen. We kunnen ons bijvoorbeeld baseren op de kwartieren om dat te doen. Als we nu een groot verschil vaststellen tussen het verband gemeten op basis van de ordinale maat gamma en de originele correlatiecoëfficiënt, dan is dat een teken dat er iets aan de hand is wat nadere inspectie vereist. Het kan betekenen dat de associatie niet rechtlijnig is. De beste manier om dat na te gaan is om een puntenwolk op te vragen.

## **8. De rangcorrelatiecoëfficiënt van Spearman en Kendall's Tau-*b***

Wanneer onze data gemeten zijn op ordinaal niveau en de waarnemingen van onze studie kunnen gerangschikt worden in twee onderscheiden reeksen, kunnen we gebruik maken van **rangcorrelatiecoëfficiënten**. In dit geval toetsen we de nulhypothese dat beide rangordeningen niet van elkaar verschillen. Wat wordt hiermee bedoeld? We kunnen de respondenten rangordenen van laag naar hoog op basis van de antwoorden op de variabele “regels zijn gemaakt om te breken”. We kunnen hetzelfde doen voor de variabele “spijbelen van vrienden”. Vervolgens kunnen we ons de vraag stellen of deze beide ordeningen samenhangen: iemand die hoog gerangschikt is op variabele X, is die ook hoger gerangschikt op variabele Y? De meest bekende rangcorrelatiecoëfficiënten zijn deze van Spearman (*Rho* genoemd) en deze van Kendall (*Tau-b*). We nemen opnieuw dezelfde associatie tussen twee

enquêtevragen over spijbelen van vrienden en de attitude tegenover regelovertraving als voorbeeld.

**Kruistabel: regels zijn gemaakt om te breken en spijbelen van vrienden**

		Regels zijn gemaakt om te breken					Totaal
		helemaal oneens	beetje oneens	noch eens, noch oneens	beetje eens	helemaal eens	
<b>Spijbelen van vrienden</b>	<b>Geen vriend</b>	811	261	265	185	94	1616
	<b>Eén vriend</b>	221	154	159	130	83	747
	<b>Twoe vrienden</b>	11	10	15	16	15	67
	<b>Drie of meer vrienden</b>	9	5	9	4	11	38
<b>Totaal</b>		1052	430	448	335	203	2468

De rangcorrelatiecoëfficiënt van Spearman is afgeleid van de **product-moment correlatiecoëfficiënt van Pearson**. Deze coëfficiënt varieert van min één tot plus één. Nul wijst op de afwezigheid van een verband. Min één wijst op een perfect negatieve associatie en plus één wijst op een perfect positieve associatie. Om de rangcorrelatiecoëfficiënt te berekenen dienen we de respondenten die we bevraagd hebben te gaan rangordenen op deze beide kenmerken. Om **Rho** te berekenen bepalen we per eenheid het verschil tussen de beide rangordeningen. De achterliggende idee is dat samenhang tussen de beide reeksen perfect zal zijn indien de rangordeningen niet van elkaar verschillen, en dat wanneer dit laatste wel het geval is, de samenhang minder sterk zal zijn. **Kendall's Tau-b** is eveneens een symmetrische associatiemaat die varieert tussen min één en plus één. Bij perfecte statistische associatie worden de waarden plus één of min één slechts bereikt onder de conditie dat het aantal rijen even groot is als het aantal kolommen. Een analyse aan de hand van SPSS toont ons volgende resultaten.

**Tabel: Rangcorrelaties tussen twee variabelen**

		Regels zijn gemaakt om te breken
Kendall'sTau-b	spijbelen vrienden	0.203
Spearman's Rho	spijbelen vrienden	0.226

Er is dus een matige positieve associatie tussen beide ordinale kenmerken.

## 9. Leerdoelen

In dit hoofdstuk werd de opmerking gemaakt dat er een zeer belangrijk verband bestaat tussen het kiezen van een associatiemaat en het meetniveau van variabelen. Daar waar men in de univariate beschrijvende statistiek enkel en slechts enkel met het meetniveau van één kenmerk dient rekening te houden, is dit iets complexer in de bivariate statistiek. Hier dienen we rekening te houden met het meetniveau van de twee kenmerken. **Als regel geldt dat wanneer twee meetniveaus verschillen, we het laagste meetniveau kiezen.** We hebben in dit hoofdstuk een reeks van associatiematen gezien die van toepassing zijn op de studie van relaties tussen kernmerken gemeten op lagere meetniveaus. We kunnen daarbij het onderscheid tussen het ordinale en nominale meetniveau maken. We hebben gezien dat nominale associatiematen vooral gebaseerd zijn op het percentageverschil en ook op chi-kwadraat. Je kan echter twee percentageverschillen berekenen. Je moet dus goed redeneren: welke variabele beschouw je als onafhankelijke variabele en welke variabele beschouw je als afhankelijke variabele? Inhoudelijk moet je heel goed weten wat een chi-kwadraat waarde betekent. Ook de formule en de berekening van een chi-kwadraat worden verwacht gekend te zijn. We leggen ook de klemtoon op het correct interpreteren van de associatie en het trekken van een besluit, i.e. het beantwoorden van een onderzoeksvergadering waar twee variabelen bij betrokken zijn. Je dient ook de mogelijkheden en beperkingen van associatiematen voor categorische variabelen te kennen. Eveneens moet je goed het onderscheid tussen zulke associatiematen begrijpen. Je dient te weten wanneer je Phi dan wel Cramer's V gebruikt, of wanneer je een contingentietabel analyseert aan de hand van een percentageverschil. Ook dien je de odds en odds-ratio te kennen. Deze dien je zelf met behulp van een rekenmachine te kunnen uitrekenen. Tot slot dien je te weten wanneer je kiest voor gamma of Kendall's tau-b. Gamma, Rho en Kendall's tau-b moeten niet zelf kunnen berekend worden. De theoretische begrippen die werden behandeld moeten echter wel bekend zijn. Tot slot: denk steeds na wat de resultaten betekenen in functie van de onderzoeksvergadering die altijd aan de basis ligt van een statistische bivariate analyse.

## **Hoofdstuk 8**

### **Correlatie- en regressieanalyse**

#### **1. Symmetrische associatiematen voor kenmerken op metrisch niveau**

Dit hoofdstuk is een van de belangrijkste in een inleidend handboek voor toegepaste statistiek voor sociale wetenschappers. Dit hoofdstuk handelt over de lineaire samenhang tussen metrische kenmerken. Je kan je afvragen waarom deze methoden zo populair zijn in de criminologie. Je zou kunnen stellen: veel variabelen zijn toch niet echt van het metrische meetniveau? Veel onderzoek in de criminologie is gebaseerd op indicatoren die worden samengesteld op basis van vragenlijsten. Veel vragenlijsten zijn gebaseerd op uitspraken en die uitspraken hebben een beperkt aantal antwoordcategorieën. De antwoorden uit onderzoek met vragenlijsten kunnen worden verwerkt tot indicatoren die heel gedetailleerde informatie bevatten, alvast informatie die gedetailleerd genoeg is om lineaire correlatie-analyse en lineaire regressie-analyse op toe te passen. Daarnaast bestaan er natuurlijk wel een heleboel kenmerken die van nature uit metrisch zijn van aard. Een voorbeeld is het aantal delicten dat jongeren plegen binnen de tijdspanne van een jaar, het aantal maal dat men slachtoffer wordt van een misdrijf, het aantal keren dat men in de gevangenis werd opgesloten voor een druggerelateerd feit, het aantal (buitenechtelijke) kinderen dat men heeft, hoeveel maanden men leefloon geniet, enz.

Al deze kenmerken kunnen gemeten worden met een grote precisie, en kunnen met criminaliteit of de maatschappelijke reactie op criminaliteit in verband gebracht worden. Criminologen hebben zich sinds de negentiende eeuw heel intensief bezig gehouden met dergelijke vraagstellingen. De biologische school zocht naar biologische oorzaken van delinquent gedrag, de psychologische school zocht naar verbanden tussen de scores op psychologische testen en regelovertredend gedrag. Zulke studies leveren een antwoord op de vraag of er bijvoorbeeld een verband bestaat tussen de scores van iemand op een psychopathieschaal en de frequentie waarmee een persoon gewelddadige handelingen stelt. Dat zijn heel belangrijke vragen want ze leren ons iets over kenmerken zoals psychopathie en de kenmerken waarmee dit samenhangt. Vandaag de dag neemt men in de reclassering en hulpverlening vaak het geïntegreerd biospsychosociaal model, waarbij gekeken wordt naar de samenhang tussen kenmerken gemeten op individueel niveau en kenmerken gemeten op groepsniveau als uitgangspunt.

Eén van de meest volledige overzichtswerken van alle correlaties is het grote correlatiehandboek “The Handbook of Crime Correlates” van de criminologen Lee Ellis, Kevin

Beaver en John Wright. Dit handboek bevat een overzicht van kenmerken, gerangschikt volgens type en sterkte.

Om al die interessante studies te kunnen begrijpen, is het belangrijk dat je basisinzichten verkrijgt in de regressie- en correlatieanalyse. Stel je voor dat je niks kent van deze associatiematen en je gebruikt resultaten uit het eerste en beste artikelje uit een tijdschrift in het kader van een paper die je moet schrijven of erger nog: in het kader van je scriptie. De kans is reëel dat je het artikel foutief interpreteert en ook dat je niet in staat bent het kaf van het koren te onderscheiden. Het duurt wel eventjes voor je dat kan, maar Rome werd ook niet op één dag gebouwd. Zo is het ook gesteld in de criminologie-opleiding. De criminoloog wordt ook niet op een dag gevormd. Gelukkig maar. Criminologen die gedetineerden loslaten op basis van foutief geïnterpreteerde tests, zijn geen goede zaak voor de samenleving. De tijd dat je dacht dat enkel ingenieurs een goede statistische opleiding moeten hebben, is wel degelijk voorbij. Je ontsnapt dus niet aan de lineaire regressie en correlatieanalyse. Maar geloof het of niet, de correlatie- en regressieanalyse kunnen best interessant zijn eenmaal je de logica ervan doorhebt. In het tijdperk van “Big Data” waar alle informatie voor het rapen ligt, en waar zoveel digitale informatie wordt bijgehouden, kan je als (criminologische) data-analist goed aan de slag en beleef je gouden tijden. Genoeg reclamepraat echter. We komen terug tot de orde van de dag.

De lineaire samenhang tussen twee kenmerken van het metrisch niveau kan worden bestudeerd aan de hand van de **covariatie**, **de covariantie** en **de correlatiecoëfficiënt**. Deze symmetrische associatiematen zijn verwant aan elkaar. Zij vormen de basis om de bivariate regressieanalyse te begrijpen. Aan regressieanalyse wordt in een volgende paragraaf aandacht besteed. Stel dat we geïnteresseerd zijn in de samenhang tussen de criminaliteitsgraad in Gentse buurten en het werkloosheidspercentage in diezelfde Gentse buurten. We verzamelen deze metrische gegevens voor alle Gentse buurten. Als blijkt dat hoge criminaliteitsgraden in buurten samenhangen met hoge werkloosheidspercentages, dan is er sprake van een positieve samenhang. Omgekeerd, als blijkt dat hoge criminaliteitsgraden samenhangen met lage werkloosheidsgraden, dan is er sprake van negatieve samenhang. Als er geen samenhang bestaat tussen beide metrische kenmerken dan zal de covariatie, covariantie maar ook de correlatie nul bedragen.

Om de bivariate lineaire samenhang tussen twee metrische kenmerken beter te begrijpen, kunnen we best beroep doen op een puntenwolk of **scatterplot**. Dit principe hebben we eerder al eens kort uitgelegd. Elk punt represeneert een statistische eenheid (bijvoorbeeld een individu, een buurt) en elk punt bevat informatie over een X-variabele en een Y-variabele. Elk

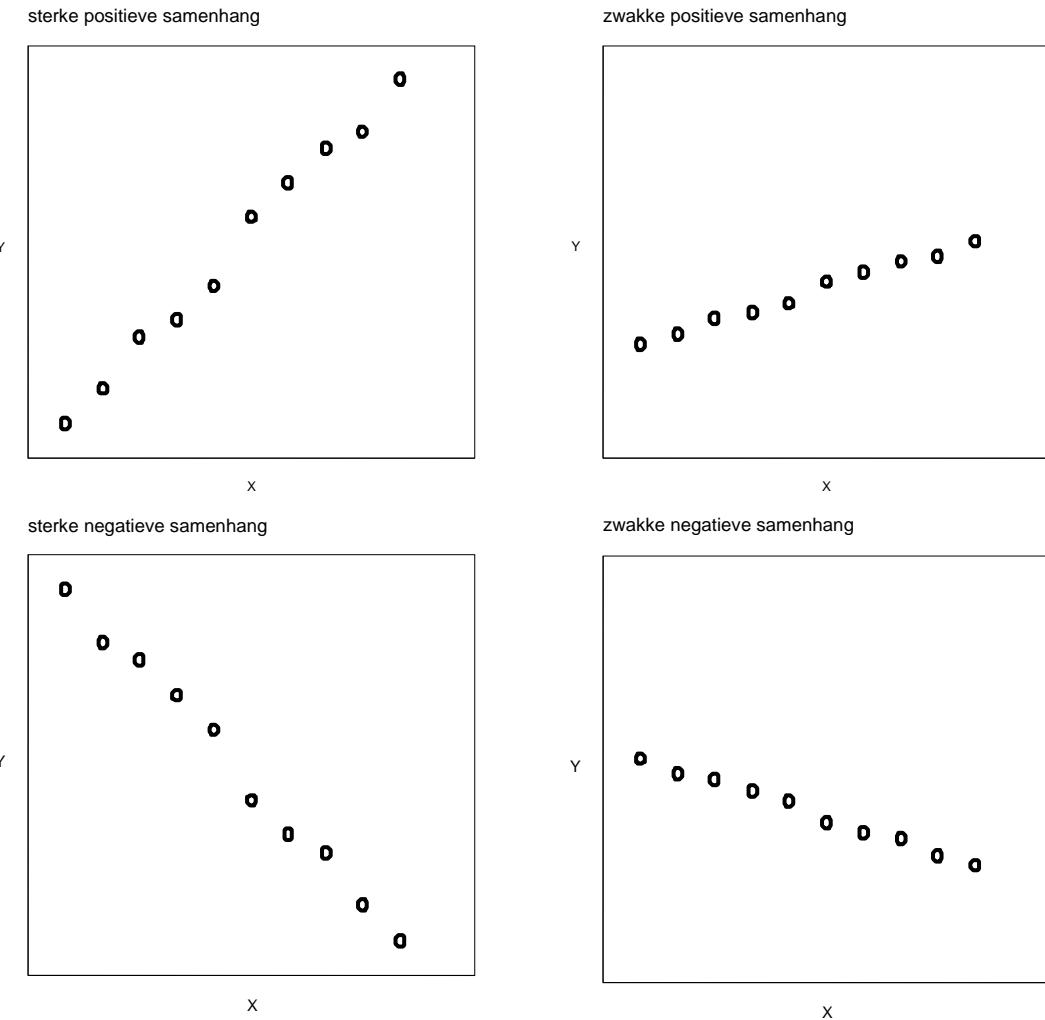
punt heeft dus wat we noemen xy-coördinaten. Deze coördinaten worden gegeven voor de kenmerken waarin men geïnteresseerd is, zoals de criminaliteitsgraad en de werkloosheidsgraad. Alle punten vormen samen een puntenwolk.

De puntenwolk is dus de verzameling van alle elementen uit onze steekproef waarbij geldt dat we voor elk element de waarde op de X-variabele en de waarde op de Y-variabele kunnen aflezen.

In de tekening hieronder zien we vier verschillende situaties. We bespreken ze even.

- Links bovenaan kan een puntenwolk gezien worden, waaruit een heel sterk positief verband blijkt te bestaan. Hoge waarden op de X-variabele gaan samen met hoge waarden op de Y-variabele. We spreken van een sterke positieve samenhang.
- Rechts bovenaan zien we eveneens een puntenwolk, maar de relatie tussen X en Y is toch minder uitgesproken. Je moet al veel beter je best doen om daar een sterk verband in te willen zien. Toegegeven, er zit een patroon in, maar de stijging is maar matig. We spreken hier duidelijk van een matig positief verband.
- Links onderaan kan een puntenwolk gezien worden, waaruit een heel sterk negatief verband blijkt te bestaan. Hoge waarden op de X-variabele gaan samen met lage waarden op de Y-variabele. We spreken van een sterke negatieve samenhang.
- Rechts onderaan zien we eveneens een puntenwolk, maar de relatie tussen X en Y is toch minder uitgesproken. Je moet al veel beter je best doen om daar een sterk verband in te willen zien. Toegegeven, er zit een patroon in, maar de daling is maar matig. We spreken hier duidelijk van een matig negatief verband.

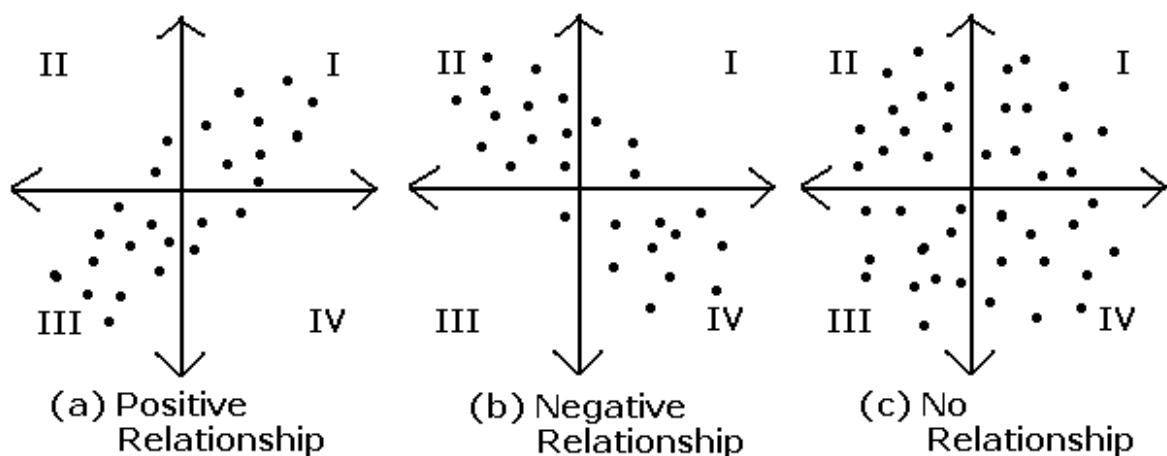
## Vormen van lineaire samenhang tussen metrische kenmerken



Omdat elk van de beide kenmerken van het metrische niveau is, kunnen we van elk punt gemakkelijk het rekenkundig gemiddelde berekenen. We kunnen een denkbeeldig punt tekenen in de puntenwolk: het punt gegeven door de gemiddelde score op de X-variabele en de gemiddelde score op de Y-variabele, noemen we **het centrale punt van de puntenwolk**. Een nog andere benaming hiervoor is het **zwaartepunt van de tweedimensionale verdeling**. Deze tweedimensionale tegenhanger van het rekenkundige gemiddelde (gedragen door de gemiddelde waarde op variabele X en de gemiddelde waarde op variabele Y) is letterlijk het

“zwaartepunt” van de verdeling: het geeft een tweedimensionale centrale waarde in het vlak (x,y). Laat ons beginnen met een illustratie.

Het bepalen van de statistische samenhang tussen twee metrische kenmerken is een indicatie van het samen optreden van afwijkingen ten opzichte van het gemiddelde bij één van de variabelen met afwijkingen ten opzichte van het gemiddelde bij de andere variabele. De richting geeft aan of het over een *positief* of een *negatief* verband gaat. Bij een positief verband hangen hoge (lage) waarden van X (Y) samen met hoge (lage) waarden van Y (X). Bij een negatief verband zullen hoge (lage) waarden van X (Y) samengaan met lage (hoge) waarden op Y (X). De *sterkte* duidt op de mate waarin beide variabelen al dan niet samenhangen. Laten we de waarnemingen in het (x,y)-vlak beschouwen en een assenkruis door het bivariate zwaartepunt tekenen.



$$\text{Kwadrant I: } (X - \bar{X})(Y - \bar{Y}) > 0$$

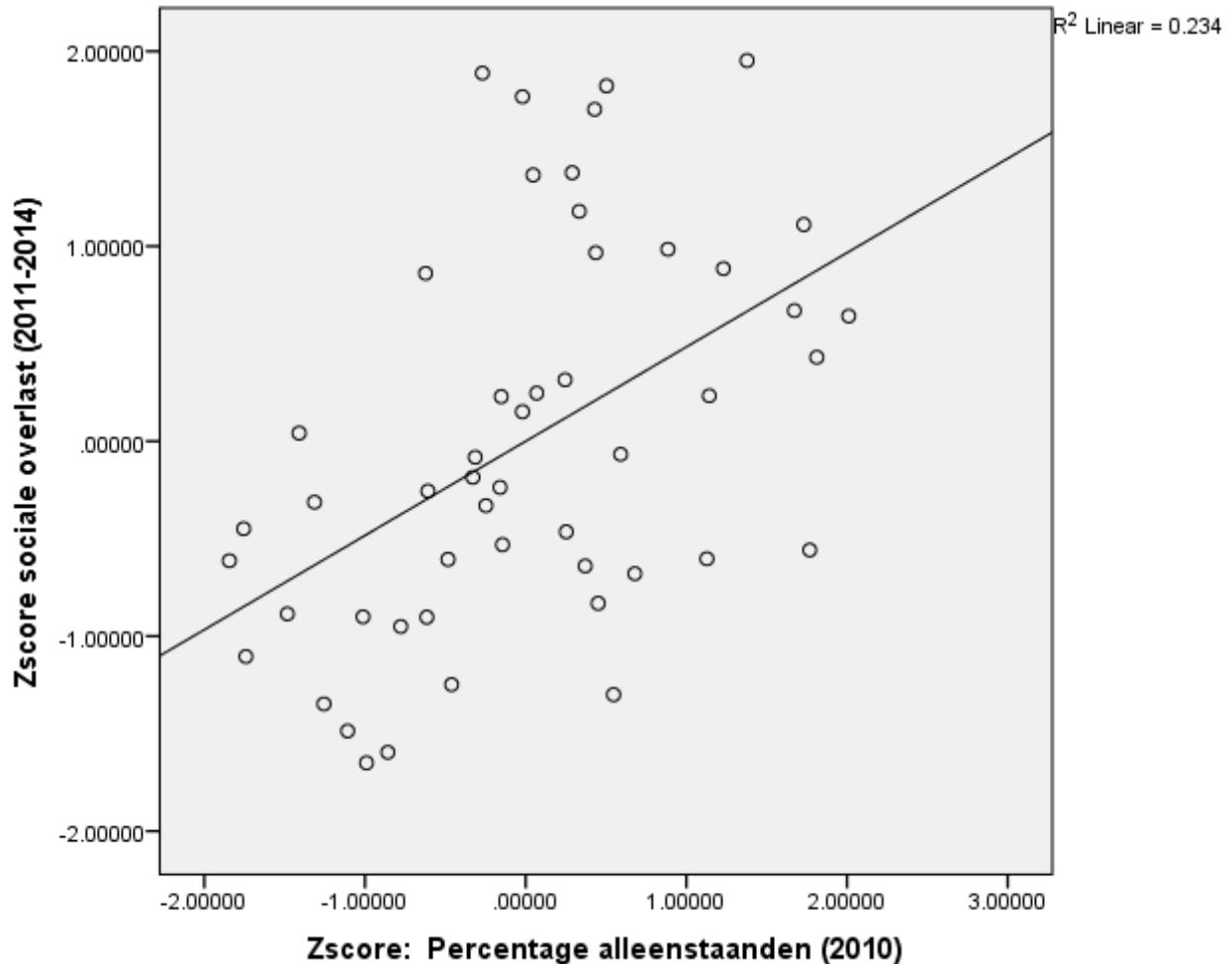
$$\text{Kwadrant II: } (X - \bar{X})(Y - \bar{Y}) < 0$$

$$\text{Kwadrant III: } (X - \bar{X})(Y - \bar{Y}) < 0$$

$$\text{Kwadrant IV: } (X - \bar{X})(Y - \bar{Y}) > 0$$

Als het merendeel der (x,y)-waarnemingen in kwadrant II en IV liggen, dan vertonen deze punten een dalende tendens. Als het merendeel der (x,y)-waarnemingen in kwadrant I en III liggen, dan vertonen deze punten een stijgende tendens.

**Voorbeeld: Puntenwolk van de relatie sociale overlast en het percentage alleenstaanden in Gentse buurten**



Laten we een voorbeeld geven uit de praktijk van het geografisch criminologisch onderzoek. Als in een puntenwolk de punten een *lineair patroon* vormen dan is er sprake van een zekere *lineaire* samenhang tussen de variabelen. In het voorbeeld hierboven is er sprake van een positieve *samenhang*. De twee variabelen worden hier gepresenteerd onder de vorm van z-scores. Dit wil zeggen dat de variabelen gestandaardiseerd zijn en dat de gemiddelde score nul bedraagt en dat de standaardafwijking één bedraagt. Als je nu naar de Y-as kijkt, dan zie je de waarde van “nul”. Idem voor de X-as. Als je deze beide punten met elkaar verbindt, dan heb je het denkbeeldige bivariate zwaartepunt. Uit deze puntenwolk leiden we een zeker patroon af: hoe hoger de score op de variabele “percentage alleenstaanden”, hoe hoger de score voor de variabele “sociale overlast”. Uit verschillende studies blijkt dat het percentage alleenstaanden

één van de sterkste predictoren voor criminaliteit en overlast lijkt te zijn. Vanuit de routine-activiteitentheorie van Marcus Felson werd gewezen op het feit dat alleenstaanden er een andere leefstijl op nahouden dan niet-alleenstaanden (althans gemiddeld genomen). Concreet komt de redenering hier op neer: in buurten waar veel alleenstaanden wonen, er minder informeel toezicht of “*guardianship*” is. Alleenstaanden gaan vaker uit dan samenwonenden en gezinnen met kinderen, waardoor de huizen en de straten aan minder toezicht worden blootgesteld. Hierdoor ontstaat een situatie waarin minder burgers bereid zijn om in te grijpen als er iets gebeurt. Dit is een interessante voedingsbodem voor criminaliteit, althans vanuit het opportunitéitsperspectief bekeken.

De hierboven geïllustreerde associatie kunnen we statistisch beschrijven aan de hand van drie belangrijke en onderling verworven associatiematen: de *covariatie*, de *covariantie* en de *correlatie*. Alvorens dieper in te gaan op de berekening van deze associatiematen, is het voor een goed begrip noodzakelijk te duiden op de mogelijkheden maar ook op de beperkingen ervan. Covariaties, covarianties en correlaties worden gebruikt om de samenhang tussen twee variabelen te schatten en het gaat hierbij om **symmetrische associatiematen**. Er is dus geen veronderstelling over causaliteit.

### ***De covariatie***

**De covariatie** wordt ook de **kruisproductensom** genoemd. De Engelstalige benaming hiervoor is de **Sum of Squares** (afgekort: **SSxy**) en ze stelt de mate voor waarin beide variabelen samen variëren (synoniem: co-variëren). Het is de **som van de kruisproducten**. Voor elke onderzoekseenheid kan je een kruisproduct berekenen. Eerder hebben we laten zien dat je voor elke onderzoekseenheid de afwijking tegenover het rekenkundig gemiddelde kan berekenen en dat vermenigvuldigen met zichzelf. Dat was eigenlijk een bijzonder geval van een kruisproduct, met name het kruisproduct met zichzelf. **Met kruisproduct bedoelen we het product** van een afwijking van een onderzoekseenheid tegenover de gemiddelde x-waarde, en de afwijking van een onderzoekseenheid tegenover de gemiddelde y-waarde. Als we deze oefening uitvoeren voor elk element in onze steekproef, dan krijgen we voor elke eenheid een nieuwe waarde. Als we die nieuwe waarden met elkaar optellen, dan krijgen we een nieuwe som. Dat is de kwadratensom en deze is heel belangrijk. Immers, op basis van de kruisproductensom worden de parameters van de bivariate associatie op metrisch niveau berekend.

$$SS(x,y) = \{ (x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \}$$

$$SS_{xy} = \sum (X - \bar{X})(Y - \bar{Y})$$

Wanneer we de berekeningswijze van de covariatie bekijken, dan zien we dat de eerder gezien variatie eigenlijk een bijzonder geval is van de covariatie. De variatie is eigenlijk *de covariatie van een kenmerk met zichzelf*. In tegenstelling tot de variatie, waar de gesommeerde deviatiescores (van één variabele) gekwadrateerd worden, vermenigvuldigt men bij de covariatie de gesommeerde deviatiescores van de twee variabelen. Deze producten noemt men kruisproducten.

### **De covariantie**

De covariatie is een maat die dezelfde nadelen heeft als de variatie. Omdat ze enkel gebaseerd is op de kruisproducten, krijgen we grote waarden. We moeten iets doen om deze maat te normeren. Een eerste belangrijke tussenstap is het berekenen van de covariantie. **De covariantie ( $S_{xy}$ ) van  $x$  en  $y$  is de *kruisproductensom* van  $(x_i - \bar{x})$  en  $(y_i - \bar{y})$ , gedeeld door  $n-1$ .**<sup>10</sup> Dus wordt de covariantie tussen  $x$  en  $y$  als volgt berekend:

$$\text{Cov}(x,y) = \{ (x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \} / n-1$$

$$S_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1}$$

Gezien de covariantie afhankelijk is van de meeteenheid waarin de variabelen zijn opgenomen kan de absolute waarde van de covariantie weinig informatie bieden over de *sterkte* van de samenhang. Door te covariatie te delen hebben we het probleem van de normering nog niet opgelost. Vandaar dat ook hier nog steeds een vaste boven- en benedengrens ontbreekt, en dus ook het ontbreken van de mogelijkheid om waarden onderling te vergelijken. Wanneer bij wijze van voorbeeld inkomen wordt uitgedrukt *in jaarlijks inkomen* en leeftijd in jaren, bekomt men een grotere waarde van de covariantie dan wanneer men werkt met *wekelijks inkomen*. Dat zou

---

<sup>10</sup> Merk op dat we ook hier delen door  $n-1$  en niet door  $n$ . De redenering is dezelfde als voorheen werd uiteengezet. **Wanneer onze gegevens berusten op steekproefresultaten wordt  $n-1$  in de noemer gebruikt en wanneer we beschikken over gegevens afkomstig uit een voltallige populatie dan delen we door  $n$ .**

ook weer niet mogen. Een associatie tussen twee kenmerken mag niet afhankelijk zijn van de meting (uitgedrukt in weken dan wel in jaren). Een grotere waarde van de covariantie duidt dus niet op een sterkere samenhang maar is een rechtstreeks gevolg van het feit dat de numerieke waarde en de spreiding van inkomen groter is bij jaarlijks dan bij wekelijks inkomen. Aan de relatie tussen inkomen en leeftijd is echter niets veranderd. De oplossing is het standaardiseren van de covariantie. Deze gestandaardiseerde covariantie is gekend als de product-moment correlatiecoëfficiënt van Pearson.

### ***De product-moment correlatiecoëfficiënt van Pearson***

De **correlatiecoëfficiënt**, ook wel product-moment correlatiecoëfficiënt van Pearson genoemd, is gelijk aan de covariantie tussen X en Y in gestandaardiseerde vorm. De formule ziet er als volgt uit:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Om de correlatie te berekenen volstaat het dus de covariantie Cov(x,y) te standaardiseren door de covariantie te delen door  $S_x \cdot S_y$ . Standaardisatie maakt de interpretatie van de associatie tussen twee metrische variabelen eenvoudiger. De correlatie varieert dankzij het proces van standaardisering van min één tot plus één; waarbij nul wijst op gebrek aan correlatie.

### **Hier volgen enkele vuistregels voor de interpretatie van de associaties:**

- **0 – 0,10 zeer zwak/geen verband;**
- **0,11 – 0,30 zwak verband;**
- **0,31 – 0,50 redelijk verband;**
- **0,51 – 0,80 sterk verband;**
- **0,81 – 0,99 zeer sterk verband;**
- **1 perfect verband.**

Het gaat hier om absolute waarden, ongeacht het teken. Er dient verder op gewezen te worden dat men veelvuldig de fout maakt om op basis van een lage waarde van de correlatie tussen twee variabelen te besluiten dat er geen verband zou bestaan tussen beide kenmerken. Men dient zich steeds te realiseren dat een correlatie toetst naar een *lineair* verband tussen twee variabelen. Het verdient aanbeveling om een puntenwolk (of scatterplot) van beide variabelen

te bestuderen aangezien het mogelijk is dat de variabelen wel een samenhang vertonen maar dat dit verband niet lineair is.

## 2. Covariatie, covariantie en correlatie: een uitgewerkt rekenvoorbeeld

Een heleboel complexe multivariate analysetechnieken die criminoloog-onderzoekers gebruiken zijn gebaseerd op deze bivariate associatiematen. Precies omdat deze associatiematen zo belangrijk zijn in de statistiek, moeten studenten deze zeer goed kunnen interpreteren, maar ook beseffen hoe deze maten uitgerekend worden. We geven een uitgewerkt rekenvoorbeeld. We berekenen de associatie tussen twee testscores die studenten hebben op twee psychosociale proeven ter voorbereiding van een sollicitatiegesprek voor een job als strategisch analist bij de federale politie. We zetten even de stappen op een rijtje die je moet maken om de covariantie zelf uit te rekenen. In de praktijk maken criminologen echter gebruik van statistische verwerkingspakketten om deze berekening uit te voeren. Toch is het noodzakelijk dat je snapt wat er in werkelijkheid achter de schermen gebeurt wanneer je de samenhang tussen kenmerken berekent aan de hand van software pakketten. Het begrijpen van wat een statistische analyse doet, zal ervoor zorgen dat de kans dat je de resultaten van een analyse verkeerd interpreteert, aanzienlijk verkleint.

**Tabel: tussenstappen bij het berekenen van een correlatie**

Student	ScoreT1	ScoreT2	$x_1 - \bar{x}$	$y_1 - \bar{y}$	$(x_1 - \bar{x})^* (x_1 - \bar{x})$	$(y_1 - \bar{y})^* (y_1 - \bar{y})$	$(x_1 - \bar{x})^* (y_1 - \bar{y})$
An	30,00	65,00	0	2	0	4	0
Arno	45,00	75,00	15	12	225	144	180
Bart	35,00	60,00	5	-3	25	9	-15
Björn	20,00	50,00	-10	-13	100	169	130
Delphine	40,00	80,00	10	17	100	289	170
Hanne	35,00	75,00	5	12	25	144	60
Henk	30,00	70,00	0	7	0	49	0
Ines	30,00	75,00	0	12	0	144	0
Jeroen	25,00	55,00	-5	-8	25	64	40
Jurgen	20,00	40,00	-10	-23	100	529	230
Kim	40,00	75,00	10	12	100	144	120
Robert	25,00	60,00	-5	-3	25	9	15
Nele	20,00	60,00	-10	-3	100	9	30
Sara	25,00	50,00	-5	-13	25	169	65
Sofie	30,00	55,00	0	-8	0	64	0
N= 15					Sum of squares (Variatie) 850	Sum of squares (Variatie) 1940	Covariatie of Kruisproductensom 1025
$\bar{x} = 30$							
$\bar{y} = 63$							

### ***Stappen te volgen in het uitrekenen van een correlatie:***

Hieronder hebben we de stappen uitgeschreven. We brengen in herinnering dat we gebruik maken van dezelfde gegevens, met name de testscores van studenten op twee examens. Echter, in plaats van te kijken naar de individuele variatie in de testscores, gaan we nu kijken naar de mate waarin deze twee testscores samenhangen.

1. Bereken het ***rekenkundig gemiddelde*** van de twee variabelen, zoals eerder werd uiteengezet.
2. Bereken de ***afwijkingen*** van elke onderzoekseenheid ten opzichte van het rekenkundig gemiddelde voor de beide variabelen X en Y. Anders gesteld: bereken de deviatiescores voor elke onderzoekseenheid op basis van de variabelen X en Y.
3. ***Kwadrateer de deviatiescore op basis van X en op basis van Y.*** Op die manier leg je de basis voor de berekening van de variatie in X en Y en de covariatie tussen X en Y.
4. Neem de ***kwadratensom van de deviatiescores op basis van X en op basis van Y.***
5. Bereken de ***variantie van X*** en de ***variantie van Y***. Dit gebeurt door zowel variatie van X als de variatie van Y te delen door N-1.
6. Bereken de vierkantswortel van de variantie in X en de vierkantswortel van de variantie in Y. Je hebt nu ook de ***standaardafwijking*** van X en van Y.
7. Bereken de ***kruisproductensom*** en je hebt de ***covariatie tussen X en Y.***
8. ***Deel de kruisproductensom door n-1*** en je hebt de ***covariantie*** tussen X en Y.
9. ***Vermenigvuldig*** de standaardafwijking van X met de standaardafwijking van Y.
10. Deel de covariantie tussen X en Y door de vermenigvuldiging van de standaardafwijking van X met de standaardafwijking van Y. Dit resultaat is de ***correlatiecoëfficiënt van Pearson.***

Als je berekening juist is stel je vast dat de correlatie 0.79 bedraagt. Een heranalyse met het statistisch verwerkingspakket SPSS leert ons hetzelfde. De uitkomst van deze heranalyse ziet er zo uit:

### Descriptive Statistics

	Mean	Std. Deviation	N
Score T1	30,0000	7,79194	15
Score T2	63,0000	11,77164	15

### Correlations

		ScoreT1	ScoreT2
Score T1	Pearson Correlation	1	,798(**)
	Sum of Squares and Cross-products	850,000	1025,000
	Covariance	60,714	73,214
	N	15	15
ScoreT2	Pearson Correlation	,798(**)	1
	Sum of Squares and Cross-products	1025,000	1940,000
	Covariance	73,214	138,571
	N	15	15

\*\* Correlation is significant at the 0.01 level (2-tailed).

### 3. De bivariate lineaire regressieanalyse als asymmetrische analysetechniek

Ook wetenschappers en beleidsmakers houden van voorspellingen dat iemand een misdrijf gaat plegen. Zou het mogelijk zijn om te interveniëren vlak voor het misdrijf plaatsvindt? Neen, we leven nog steeds niet in zo een “brave new world”. Denk bijvoorbeeld aan de film “Minority Report”. Maar toch is voorspellen nuttig. Rechters willen graag rekening houden met recidivestudies en criminale carrièrestudies wanneer zij de moeilijke beslissingen moeten nemen wat te doen met iemand die voor de rechter verschijnt. Een jonge student verschijnt voor de eerste keer voor een rechter. Deze jongeman heeft te snel gereden. Deze jongeman heeft geen criminale antecedenten en goede schoolresultaten en bovendien alles om het te maken. Het kan zijn dat deze jongeman eventjes is uitgeglezen en een fout gemaakt heeft. Dat gebeurt. Een andere jongeman heeft een ernstig coke-probleem en heeft onder invloed gereden. De kans dat die laatste nog eens onder invloed achter het stuur kruipt is mogelijk groter dan de kans dat de andere jongeman met het onberispelijke verleden onder invloed achter het stuur kruipt. Zou het mogelijk zijn te voorspellen op basis van een reeks achtergrondkenmerken wat er gebeurt? Stel je nu eens voor dat je in het kader van je masterproef geïnteresseerd bent in criminale recidive. Je wil wel eens weten of je kan voorspellen hoeveel keer iemand in de gevangenis zal belanden in de volwassenheid op basis van de scores op een IQ-test die werd afgenoemt in de kindertijd. Een eerste vraag die bij je opkomt is wellicht: kan dat wel? Immers, je verzamelt informatie over individuen in een steekproef als ze pakweg allemaal tien jaar zijn en na dertig jaar ga je deze steekproefpersonen terug opzoeken en je stelt hen bijkomende vragen en vraagt

hun criminale gegevens op. Er bestaan studies die dergelijke gegevens hebben bijgehouden en er zijn criminologen die dat onderzocht hebben. Eén van die historische figuren was professor David Farrington. Als je voorspellingen wil maken heb je meestal wel wat criminologische theorie als gods. Denk aan wat in het inleidende hoofdstuk werd gezegd. We gaan gemakshalve uit van de idee dat er een verband bestaat tussen IQ en criminale recidive. Waarom zou zo een verband er moeten zijn? Wel, het is mogelijk dat individuen met een laag IQ minder goed de gevolgen van hun gedrag inschatten, of minder goed kunnen inschatten of ze wel eens door de politie zouden opgepakt kunnen worden. Dus: onze hypothese is zeker plausibel. Ze is zelfs onderzocht door de criminologen James Q. Wilson en Richard Herrnstein in de jaren 1980. We verzamelen gegevens over de criminale recidive, die we onze afhankelijke variabele noemen. Deze afhankelijke variabele noemen we ook wel het explanandum. We willen weten of we de scores op de afhankelijke variabele kunnen verklaren op basis van een onafhankelijke variabele. De onafhankelijke variabele is het explanans. We gaan een voorspelling maken en we doen dat door een lineaire regressieanalyse uit te voeren.

De volledige naam van deze analysetechniek is de *OLS-lineaire regressieanalyse*. OLS staat voor *Ordinary Least Squares* en is het principe dat gebruikt wordt om de regressieanalyse mathematisch uit te voeren. Het begrip OLS leggen we verderop uit. Hou dat begrip nog even in het achterhoofd. De regressieanalyse is eerst en vooral een **asymmetrische** associatiemaat. We begeven ons nu op het terrein van de verklarende statistiek, die gebruikt wordt ter toetsing van criminologische theorieën en die gebruikt wordt in het predictie-onderzoek. Het doel van een bivariate lineaire regressieanalyse bestaat uit het statistisch verklaren van de variatie in een **afhankelijke variabele** (de responsvariabele of het explanandum genoemd) op basis van een **onafhankelijke variabele** (predictor-variabele of explanans). Er wordt in het theorietoetsend onderzoek een causaal verband tussen beide variabelen verondersteld waarmee men de afhankelijke variabele tracht te verklaren. In het predictie-onderzoek wordt dat causaal verband niet altijd verondersteld, men wil vooral weten of een variabele een voorspellend karakter heeft, want men wil bij de berechting rekening houden met een reeks van kenmerken van de persoon in kwestie.

Het uitvoeren van een enkelvoudige lineaire regressieanalyse levert een statistische vergelijking op waarmee de afhankelijke variabele voorspeld kan worden op basis van de onafhankelijke variabele. Bij enkelvoudige lineaire regressie kan deze vergelijking op twee manieren genoteerd worden:

$$Y = \beta_0 + \beta_1 X + e \quad \text{of} \quad \text{Aantal veroordelingen} = \beta_0 + \beta_1 IQ + e$$

of nog

$$Y = a + b_1 X + e$$

In deze vergelijking is **Y de geobserveerde afhankelijke variabele**. X is de onafhankelijke variabele. De  $\beta$ 's zijn de (populatie)parameters die met de regressieanalyse worden geschat en worden daarom ook wel de *regressiecoëfficiënten* genoemd.

- (1)  $\beta_0$  is de (*regressie*)*constante of het intercept*. Deze wordt ook met het symbool a aangeduid. Dat zagen we in de tweede notatie. **Deze constante drukt de verwachte of voorspelde waarde van Y uit wanneer X nul bedraagt.**<sup>11</sup> Het intercept is niet steeds betekenisvol. Onderzoekers zijn veel meer geïnteresseerd in de effecten van de onafhankelijke variabele.
- (2)  $\beta_1$  (in de tweede notatie  $b_1$ ) is het ongestandaardiseerde *regressiegewicht* dat de helling van de regressielijn aanduidt.  **$\beta_1$  geeft aan met hoeveel eenheden Y toeneemt als X met één eenheid toeneemt.** Als het IQ van een persoon gemeten is aan de hand van een gekende IQ test (zoals bijvoorbeeld de Stanford binet test) dan is de richtingscoëfficient gelijk aan de toename in Y als het IQ met een eenheid toeneemt. Dat zegt wellicht ook niet veel. Daarom kunnen we een onafhankelijke variabele ook standaardiseren vooraleer we de analyse uitvoeren. Dan krijgt de richtingscoëfficiënt wel betekenis. Een standaardafwijking heeft betekenis. We weten dat IQ normaal verdeeld is en we kunnen ons dus wel iets voorstellen bij een toename van een standaardafwijking. De waarde van het regressiegewicht kan zowel positief als negatief zijn. Bij een positieve waarde treedt er een stijging op van Y per eenheidstoename van X. Bij een negatieve waarde treedt er een daling op van Y per eenheidstoename van X. Hoe groter de waarde van  $\beta_1$  des te groter de verandering van Y bij een verandering van X. Door te werken met een gestandaardiseerde onafhankelijke

---

<sup>11</sup> Wil men dat deze waarde wel informatief wordt, kan men de onafhankelijke variabele centreren rond het gemiddelde. Op die manier staat nul voor een gemiddelde waarde, en krijgt het intercept de betekenis van de score op Y voor iemand met een gemiddelde score op X.

variabele krijgt ook het intercept betekenis, want dit is de waarde voor Y als X nul is. We weten dat nul wijst op de gemiddelde score, als X gestandaardiseerd is.

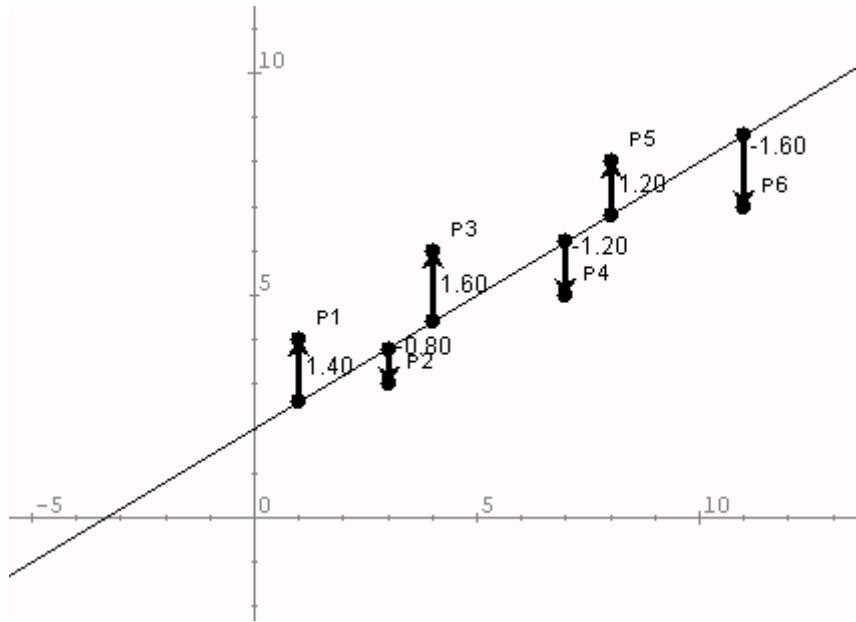
- (3) De  $e$  in de vergelijking staat voor de **foutenterm** ('error'). Dit is het **verschil tussen de werkelijke (geobserveerde) waarde van Y en de door het statistische regressiemodel voorspelde waarde van Y**. Geen enkele voorspelling is immers perfect in de sociale wetenschappen. Er is in de praktijk dus altijd een verschil tussen een observatie en een predictie. Voor sommige individuen zullen we heel goede voorspellingen kunnen maken, voor anderen heel slechte voorspellingen. Dit verschil tussen de predictie en de geobserveerde waarde wordt ook wel het *residu of de residuele term* genoemd. De som van alle residuen bedraagt nul. De variantie van  $e$  is gelijk ongeacht de waarde van X. Deze foutenterm is niet gecorreleerd met X.

Op onderstaande grafiek zijn de punten  $P_1(x_1, y_1)$  tot en met  $P_6(x_6, y_6)$  van een puntenwolk getekend. De rechte  $y = a + bx$  werd ook getekend en de punten  $(x_i, \hat{y}_i)$  op de rechte die telkens corresponderen met de punten  $P_i$ .

*Opmerking: de voorspelde waarde op basis van X wordt aangeduid met de notatie  $\hat{y}_i$*

*We spreken  $\hat{y}_i$  uit (Engels: "Y-hat") als de verwachte / voorspelde waarde van Y op basis van X.*

**Figuur: De regressierechte van y gegeven x**



Bekijk de relatie tussen x en y in de figuur hierboven. Indien men veronderstelt dat y afhangt van x, bepaalt men de regressierechte van y gegeven x. Deze lijn vat de lineaire relatie tussen beide variabelen zo goed mogelijk samen. Een algemene voorstelling van een lineair verband tussen twee variabelen wordt gegeven door de formule. Men beschouwt een theoretisch model in hetwelk de veranderlijke y kan beschouwd worden als een lineaire functie van x

$$Y = \beta_0 + \beta_1 X + e$$

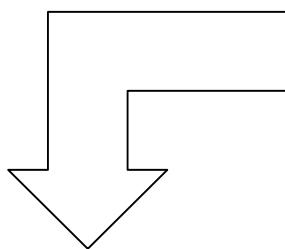
Voor elke waarneming  $x_i$  heeft men twee overeenkomende waarden van y: deze zijn: (1) de geobserveerde waarneming  $y_i$  en (2) de voorspelde waarde op basis van de onafhankelijke variabele, waarbij men uitgaat van lineariteit.

Het is ontzettend belangrijk de redenering en de betekenis van de parameters te kennen, zodat men geen foutieve interpretatie maakt van de resultaten van een regressieanalyse. Men dient de geschatte regressieparameters goed te kunnen interpreteren en verwoorden in termen van inhoudelijke betekenis.

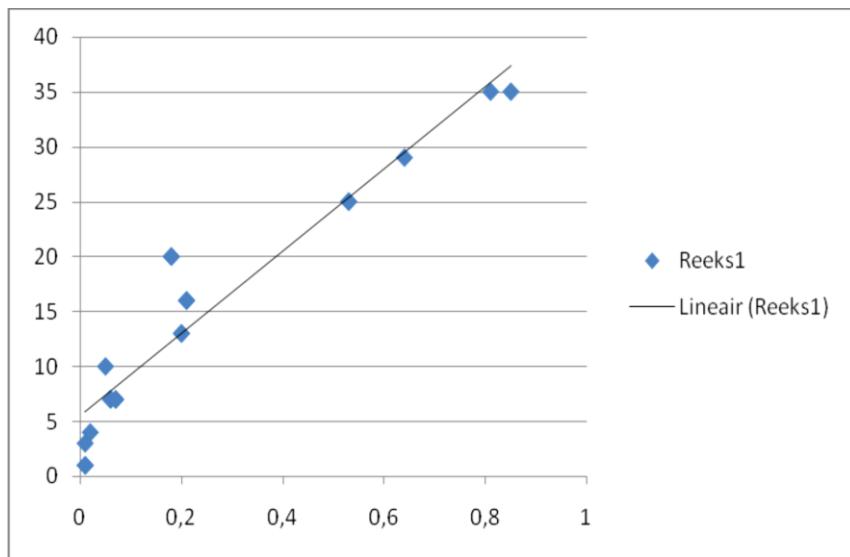
Als de geobserveerde waarden van de afhankelijke variabele in een puntenwolk geplaatst worden tegenover de geobserveerde waarden van de onafhankelijke variabele krijgt men een visueel beeld van (x,y)-coördinaten. Nu tracht men via een regressievergelijking het verband

tussen de onafhankelijke en afhankelijke variabele voor te stellen door middel van een rechte zodanig dat de globale voorspellingsfout die men maakt zo klein mogelijk is. We trachten de analysetechniek zo eenvoudig mogelijk voor te stellen. Hieronder zien we voor tien statistische eenheden de observaties voor twee criminologisch relevante kenmerken, met name X en Y. Uit de ruwe observaties is het niet duidelijk of deze kenmerken samenhangen. We maken een puntenwolk en we zien een sterke samenhang. Het lijkt er sterk op dat we een voorspelling kunnen maken van Y op basis van X. Dit zien we alleen al uit de puntenwolk. Maar hoe goed zijn deze voorspellingen? En hoe drukken we deze uit? De bivariate regressieanalyse is de techniek bij uitstek die je deze onderzoeksraag leert te beantwoorden

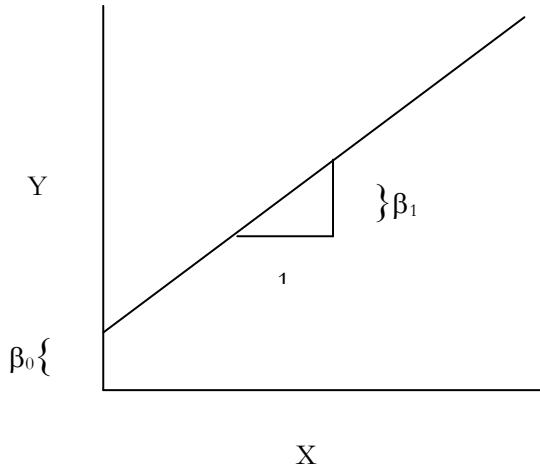
Figuur: X-Waarden en Y-waarden voor tien statistische eenheden



Eenheid	X	Y
1	0,64	29
2	0,21	16
3	0,85	35
4	0,53	25
5	0,02	4
6	0,01	1
7	0,21	16
8	0,18	20
9	0,06	7
10	0,20	13



**Hoe bepalen we het intercept en de hellingshoek (richtingscoëfficiënt)?**



**Het bepalen van de regressiecoëfficiënten gebeurt via hetgeen men noemt “de gewone kleinste kwadratenoplossing”.** Om een zo goed mogelijk model te bekomen, worden de parameters  $a$  en  $b$  zodanig bepaald dat de gekwadrateerde afwijking van de regressierechte minimaal wordt. Anders gezegd: de gewone kleinste kwadraten methode (Engels: *Ordinary*

**Least Squares of OLS**) levert een zodanige formule voor de regressielijn dat de gekwadrateerde afstand van alle datapunten tot die lijn (Engels: **residual sum of squares**) minimaal is. Bijgevolg is de kwadratische afwijking van de punten t.o.v. de regressierechte een kwaliteitsmaat voor de gevonden rechte. Men noemt deze de *residuele variatie*. De regressielijn gaat per definitie door het zwaartepunt van de puntenwolk. **Het minimaliseren van de residuen (verschillen tussen observaties en voorspellingen) is wat er gebeurt in deze asymmetrische analysetechniek.**

We stellen nu de regressievergelijking op en hanteren het principe van de kleinste kwadratenoplossing om de onderzoeksvraag te beantwoorden:

$$\hat{y}_i = a + b_1 X_i$$

In deze vergelijking zijn  $a$  en  $b_1$  de steekproefparameters;  $a$  is *de regressieconstante* en  $b_1$  is het *regressiegewicht*.  $\hat{y}_i$  is dus de y-waarde die voorspeld wordt voor subject  $i$  op basis van de waarde van  $x$  voor subject  $i$ . Nu is het dus de bedoeling  $a$  en  $b_1$  zo te kiezen dat de som van de gekwadrateerde afwijkingen tussen de geobserveerde waarden en de verwachte waarden zo klein mogelijk is. Er kan worden aangetoond dat de optimale waarden van  $b_0$  en  $b_1$  de volgende zijn:

$$a = \bar{Y} - b_1 \bar{X}$$

$$b_1 = r(X, Y) S_y / S_x$$

waarbij:

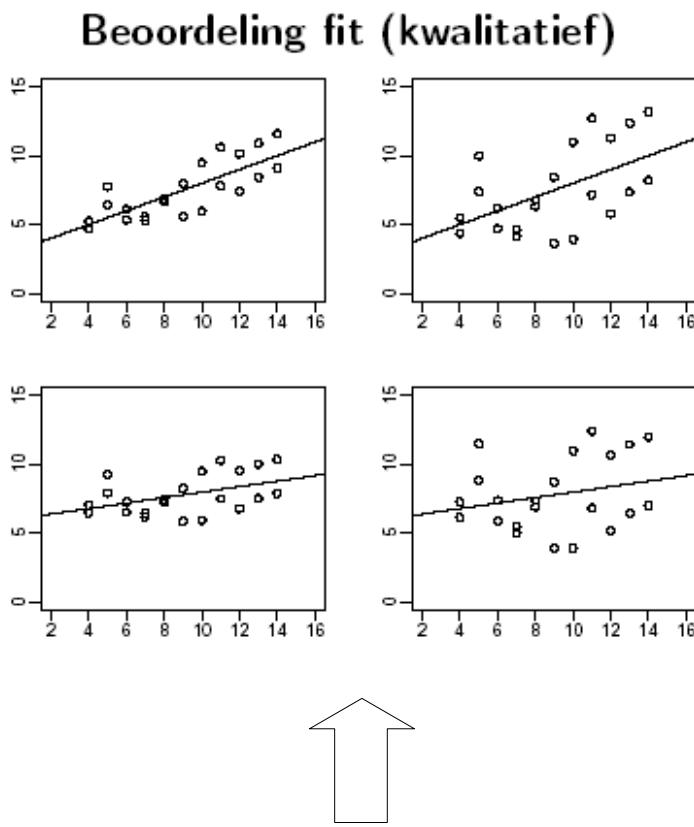
- $\bar{Y}$  en  $\bar{X}$  de (steekproef)gemiddelden zijn van  $Y$  en  $X$ ,
- $S_y$  en  $S_x$  de steekproefstandaarddeviaties
- en  $r(X, Y)$  de steekproefcorrelatie tussen  $X$  en  $Y$ .

De interpretatie van deze parameters kan gebeuren zoals hierboven werd beschreven bij de modelformulering. Bovenstaande vergelijking wordt ook wel een **statistisch model** genoemd. Een belangrijke vraag die de criminoloog dient te beantwoorden is de volgende: **hoe goed past het statistische model bij de geobserveerde data? We noemen het beantwoorden van deze vraag ook wel de evaluatie van de “model fit (Engels: to fit= passen)”**. **De model fit zegt**

**dus iets over hoe goed onze predicties het doen tegenover onze observaties. Of anders: past het statistische model goed bij de data? Als het er goed bij past, vat het de relatie tussen X en Y heel goed samen.**

We kunnen dit “tentatief” doen en een eerste blik werpen op regressiemodellen. Hieronder presenteren we een aantal situaties die zich in de praktijk van de statistische analyse kunnen voordoen. Let op het volgende: de voorbeelden werden zo geselecteerd dat je het verschil goed merkt. Er zijn situaties waarin de observaties of punten heel dicht bij de best passende lijn liggen en er zijn situaties waar de punten iets verder van de lijn liggen. Deze twee situaties zijn denkbeeldig voor elke rechte (ongeacht hoe steil deze is). Door een puntenwolk te tekenen en de best passende rechte op de puntenwolk te tekenen, zie je hoe dicht de observaties bij de lijn liggen en daardoor zie je al op het eerste zicht of de voorspelling goed dan wel zwak is. Hoe dichter bij de lijn, hoe beter.

**Figuur: model fit beoordeling op kwalitatieve wijze**



- Punten dicht bij de lijn: betere fit
- Lijn heeft grotere helling (t.o.v. range van  $Y$ ): betere fit
- In welke plaatjes dus een goede fit?

Het mag duidelijk zijn dat we met het blote oog niet in staat zijn *precieze uitspraken* te doen over de model fit. Daarom doen we beroep op een reeks coëfficiënten die de kwaliteit van de regressielijn helder uitdrukken. Eenmaal de regressiecoëfficiënten gekend zijn, dient nog bepaald te worden of en hoe goed men de waarden van de afhankelijke variabele ( $Y$ ) kan voorspellen op basis van de waarden van de onafhankelijke variabele ( $X$ ). We willen weten hoeveel van de variatie<sup>12</sup> in  $Y$  men kan “verklaren” op basis van de variatie in  $X$ .

---

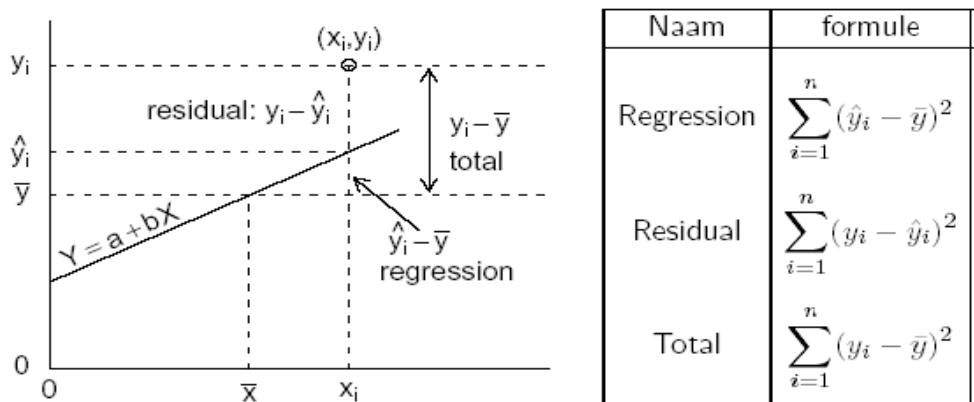
<sup>12</sup> Inhoudelijk: de verschillen in  $Y$ -waarden, maar ook statistisch: de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde in  $Y$  vormen de statistische uitdrukking voor variatie.

De afwijkingen van het gemiddelde van Y bestaan uit twee componenten: **toevallige afwijkingen (residual)** en **verklaarde afwijkingen**: afwijkingen van het rekenkundig gemiddelde verklaard door de regressielijn.

De totale variatie van Y kan opgesplitst worden in twee delen.

- **Een eerste gedeelte** vertegenwoordigt de variatie in Y die ‘verklaard’ wordt door X. We noemen dit gedeelte van de variatie in Y de “**regression sum of squares**”. Dit gedeelte wordt steeds groter naarmate men er beter in slaagt Y-waarden te voorspellen die verder afwijken van de voorspelling die men zou maken als men geen informatie heeft. Als we geen informatie hebben over een onafhankelijke variabele, dan is onze beste voorspelling nog steeds het rekenkundig gemiddelde ( $\bar{Y}$ )<sup>13</sup>.
- **Een tweede gedeelte** vertegenwoordigt de *foutenvariatie*, dit wil zeggen de mate waarin de voorgespelde waarden van Y afwijken van de geobserveerde waarden van Y. We noemen deze ook de “**residual sum of squares**”.
- De **totale variatie in Y** noemen we de *Total sum of squares* (zie univariate statistiek) en is dus mathematisch niets anders dan **de som van de Regression sum of squares en de residual sum of squares**. In formule en getekend krijgen we het volgende resultaat:

**Figuur: de regressierechte in technische termen ontleed**



Een goede maat voor de kwaliteit van de regressievoorspelling is de verhouding van de verklaarde variatie t.o.v. de totale variatie. Deze wordt de *determinatiecoëfficiënt* ( $R^2$ )

<sup>13</sup> Indien men over geen enkele informatie beschikt om de voorspelling op te baseren, maakt men de kleinste globale voorspellingsfout door het gemiddelde van de variabele als voorgespelde waarde te gebruiken.

genoemd. De coëfficiënt **R**<sup>2</sup> (“r-kwadraat”) is de proportionele reductie in de voorspellingsfout die ontstaat door op de regressielijn te steunen bij de voorspelling van de afhankelijke variabele en is dus **de proportie van de totale variatie in Y die door X wordt verklaard**. Deze coëfficiënt kan waarden aannemen gaande van 0 tot 1. In een bivariate regressieanalyse is de **determinatiecoëfficiënt gelijk aan het kwadraat van de correlatiecoëfficiënt tussen X en Y**.

Laten we de teller en noemer in de formule van de determinatiecoëfficiënt ontleden:

$$R^2 = \frac{\sum_{i=1}^n [\hat{Y}_i - \bar{Y}]^2}{\sum_{i=1}^n [Y_i - \bar{Y}]^2}$$

← Teller: Regression sum of squares  
 ← Noemer: Total sum of squares

De **teller** bevat variatie die bestaat uit de som van het gekwadrateerde verschil tussen de voorspelde waarde van Y en de gemiddelde waarde van Y. De gemiddelde waarde is immers onze beste voorspeller als we geen onafhankelijke variabele hebben.

De **noemer** bestaat uit de som van het gekwadrateerde verschil tussen de geobserveerde waarde van Y en het gemiddelde van Y.

Omdat er wordt gesteund op de regressielijn voor de voorspelling van Y, is **R**<sup>2</sup> in zuiver technisch opzicht een symmetrische maat. Het is wel mogelijk om een **asymmetrische interpretatie** aan **R**<sup>2</sup> geven.

#### 4. Zelf uitrekenen van de parameters van de regressierechte

We werken een rekenvoorbeeld uit in de hieronder gepresenteerde tabel. We baseren ons op het voorbeeld dat we eerder gaven bij de berekening van de metrische associatiematen de covariatie, covariantie en correlatie. Het is duidelijk dat je al deze maten dient te kennen, wil je een regressieanalyse verstaan.

Student	ScoreT1	ScoreT2	$x_1 - \bar{x}$	$y_1 - \bar{y}$	$(x_1 - \bar{x})^*$ $(x_1 - \bar{x})$	$(y_1 - \bar{y})^*$ $(y_1 - \bar{y})$	$(x_1 - \bar{x})^* (y_1 - \bar{y})$	Predictie Y op basis van X
An	30,00	65,00	0	2	0	4	0	63
Arno	45,00	75,00	15	12	225	144	180	81
Bart	35,00	60,00	5	-3	25	9	-15	69
Björn	20,00	50,00	-10	-13	100	169	130	51
Delphine	40,00	80,00	10	17	100	289	170	75
Hanne	35,00	75,00	5	12	25	144	60	69
Henk	30,00	70,00	0	7	0	49	0	63
Ines	30,00	75,00	0	12	0	144	0	63
Jeroen	25,00	55,00	-5	-8	25	64	40	57
Jurgen	20,00	40,00	-10	-23	100	529	230	51
Kim	40,00	75,00	10	12	100	144	120	75
Robert	25,00	60,00	-5	-3	25	9	15	57
Nele	20,00	60,00	-10	-3	100	9	30	51
Sara	25,00	50,00	-5	-13	25	169	65	57
Sofie	30,00	55,00	0	-8	0	64	0	63
<b>N = 15</b>					<b>Sum of squares 850</b>	<b>Sum of squares 1940</b>	<b>Covariatie 1025</b>	
<b><math>\bar{x} = 30</math></b>								
<b><math>\bar{y} = 63</math></b>								

### Stappen te volgen in het uitrekenen van een bivariate regressie:

1. Bereken het *rekenkundig gemiddelde* van de twee variabelen, zoals eerder werd uiteengezet.
2. Bereken de *afwijkingen* van elke eenheid ten opzichte van het rekenkundig gemiddelde voor de beide variabelen X en Y.
3. *Kwadrateer de afwijkingen* van elke eenheid met het rekenkundig gemiddelde. Op die manier leg je de basis voor de berekening van de variatie in X en Y en covariatie tussen X en Y.
4. Neem de *som van de gekwadrateerde afwijkingen* tegenover de gemiddelde waarde van X en neem de som van de gekwadrateerde afwijkingen tegenover de gemiddelde waarde van Y.
5. Bereken de *variantie* van X en de variantie van Y. Dit gebeurt door zowel variatie van X als de variatie van Y te delen door n-1.
6. Bereken de vierkantswortel van de variantie in X en de vierkantswortel van de variantie in Y. Je hebt nu ook de *standaardafwijking* van X en van Y.
7. Bereken de kruisproductensom en je hebt de *covariatie*.

8. Deel de kruisproductensom door n-1 en je hebt de *covariantie* tussen X en Y.
9. *Vermenigvuldig* de standaardafwijking van X met de standaardafwijking van Y.
10. Deel de covariantie tussen X en Y door de vermenigvuldiging van de standaardafwijking van X met de standaardafwijking van Y. Nu heb je de *correlatiecoëfficiënt*. Deze is *gelijk aan de gestandaardiseerde richtingscoëfficiënt*.
11. Bereken de *ongestandaardiseerde richtingscoëfficiënt*. Deze is gelijk aan de covariantie tussen X en Y gedeeld door de variantie in X.
12. Het *intercept* (a) kan worden berekend op basis van voorgaande info:

$$a = \bar{y} - B \bar{x} = 26.824$$

13. Op basis van de regressievergelijking  $\hat{Y} = a + b.X$ , kan nu de predictie van Y berekend worden op basis van X (bijvoorbeeld als X= 30:  $26.824 + 1.206 \cdot 30 = 63$ ).
14. Bereken de *determinatiecoëfficiënt*. Deze is gelijk aan het kwadraat van de bivariate correlatiecoëfficiënt. Let wel: dit geldt enkel in het geval van bivariate regressianalyse.
15. Bereken de *aliënatiecoëfficiënt*. Deze kan gemakkelijk berekend worden:  $1 - \text{determinatiecoëfficiënt}$ . Dit is **de proportie van de totale variatie in Y die NIET door X kan verklaard worden**.

De uitkomsten van deze regressianalyse werden met SPSS uitgevoerd en kunnen hieronder geraadpleegd worden. Als deze uitkomsten ietwat verschillen met de door jullie uitgerekende uitkomsten, bedenk dan dat SPSS niet afrondt. Jullie worden wel verwacht af te ronden.

**Beschrijvende stats voor de afhankelijke en onafhankelijke variabele**

	N	Minimum	Maximum	Mean	Std. Deviation
Test_1 (X)	15	20.00	45.00	30.0000	7.79194
Test_2 (Y)	15	40.00	80.00	63.0000	11.77164
Valid N (listwise)	15				

### Correlaties en kruisproductensom

		Test_1	Test_2
Test_1	<b>Pearson Correlation</b>	1	.798**
	Sig. (2-tailed)		.000
	<b>Sum of Squares and Cross-products</b>	850.000	1025.000
	Covariance	60.714	73.214
	N	15	15
Test_2	<b>Pearson Correlation</b>	.798**	1
	Sig. (2-tailed)	.000	
	<b>Sum of Squares and Cross-products</b>	1025.000	1940.000
	Covariance	73.214	138.571
	N	15	15

\*\*. Correlation is significant at the 0.01 level (2-tailed).

### Coefficients<sup>a</sup>

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	<b>26.824</b>	7.807		3.436	.004
Test_1	<b>1.206</b>	.252	<b>.798</b>	4.778	.000

a. Dependent Variable: Test\_2 de rico bedraagt 1.206 en deze is gelijk aan 0.798

(11.77164/7.79194) oftewel de correlatiecoëfficiënt maal de standaardafwijking van Y gedeeld door de standaardafwijking van X.

### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.798 <sup>a</sup>	<b>.637</b>	.609	7.35878

a. Predictors: (Constant), Test\_1

De determinatiecoëfficiënt is het kwadraat van de correlatiecoëfficiënt en bedraagt zoals je uit de output kan afleiden 0.637. Dit wil zeggen dat 63.7% van de variabiliteit in T2 kan verklaard worden op basis van T1.

## 5. De rapportage van de belangrijkste parameters van de regressierechte in een rapport

Een belangrijke vraag waar studenten mee worstelen, is de vraag naar de rapportage van parameters uit een regressierechte in een rapport, bachelor- of masterproef. We geven een voorbeeld uit eigen verzamelde onderzoeksgegevens over criminaliteit en werkloosheid in 210 Antwerpse buurten.

Afhankelijke variabele: criminaliteitsgraad opzettelijke slagen	Ongestandaardiseerde Coëfficiënt <b>B</b>	Gestandaardiseerde coëfficiënt <b>β</b>
<b>Intercept (B0 of a)</b>	-0.045	--
<b>Onafhankelijke variabele (B1):</b> werkloosheidsgraad	<b>1.068</b>	<b>0.682</b>

Determinatiecoëfficiënt: **0.464** (Noot: vetgedrukte coëfficiënten zijn statistisch significant- zie verder)

De criminaliteitsgraad is de uitkomstvariabele en de werkloosheidsgraad is de onafhankelijke variabele (ook wel predictor-variabele of voorspeller genoemd in de literatuur). De voorspelde waarde voor de criminaliteitsgraad van geweld =  $-0.045 + 1.068$  (werkloosheidsgraad).

Op basis van de werkloosheidsgraad zijn we in staat om 46.4 procent van de variabiliteit in de criminaliteitsgraad voor opzettelijke slagen te voorspellen. 46.4 procent van de geobserveerde verschillen in de criminaliteitsgraad van Vlaamse gemeenten kan worden verklaard vanuit de werkloosheidsgraad van de gemeente. Dit is een relatief goede voorspelling, want ongeveer de helft van de variatie in de criminaliteitsgraad kan gebeuren op basis van slechts één kenmerk, de werkloosheid. Deze bevinding mag niet op zichzelf staan, maar moet de onderzoeker-criminoloog aanzetten tot het bedenken van een verklaring. Waarom is dit zo? Welk mechanisme gaat hierachter schuil? Gemeenten zijn geen causale actoren, met andere woorden, we moeten op zoek gaan naar het causale mechanisme op een lager niveau. We moeten “de black box” opendoen en kijken welke factoren mensen er toe aanzetten om meer delicten te plegen in gemeenten met een hoge werkloosheidsgraad. De vaststelling is eeuwen oud, maar als criminoloog mogen we ons niet blindstaren op de cijfertjes op zich. Dit zou resulteren in cijferfetisjisme en dat is wel het laatste dat wij met deze cursus beogen.

## 6. En wat als de meetniveaus van twee variabelen verschillend zijn?

Hierboven werden de meest klassieke bivariate beschrijvende associatiematen besproken. Van een criminoloog wordt verwacht dat hij of zij verantwoord kan kiezen tussen een reeks associatiematen. Soms zijn de meetniveaus van twee criminologisch relevante variabelen verschillend. Strikt genomen dient men zich te houden aan het meetniveau van de variabele met het laagste meetniveau. Is de afhankelijke variabele gemeten op het metrisch niveau (bijvoorbeeld het aantal criminale veroordelingen) en de onafhankelijke variabele gemeten op het nominaal niveau (bijvoorbeeld het volgen van een bepaalde behandeling of interventie), dan dient men strikt genomen een analysetechniek te kiezen op het nominaal niveau. Hiertoe dienen we de afhankelijke variabele te hercoderen naar een lager meetniveau. Hierdoor ontstaat er eigenlijk informatieverlies. In deze situatie bestaat er een veel gebruikte analysetechniek: **het vergelijken van gemiddelden**. In deze situatie worden dan de gemiddelde scores beschreven voor elke subgroep. De vergelijking van gemiddelde scores tussen groepen komt heel vaak voor en we behandelen deze techniek later wanneer we de variantieanalyse introduceren als onderdeel van de inferentiële statistiek.

Er zijn echter enkele andere vaak voorkomende situaties waar we het nog niet over gehad hebben. Wat doe je als een variabele ordinaal is en een andere variabele nominaal? In zo een situatie kiest men strikt genomen een analysetechniek op nominaal niveau. Is een afhankelijke variabele van het nominaal niveau en een onafhankelijke variabele van het metrisch niveau, dan kiest men ook strikt genomen voor een analysetechniek gemeten van het nominaal niveau. Je merkt dat er heel wat situaties bestaan die niet voldoen aan hetgeen we tot hiertoe behandeld hebben. Dat wil zeggen dat we hier eigenlijk nog maar het topje van de ijsberg hebben behandeld. Desalniettemin is het zo dat wie zich doorheen deze cursus sleept, een goede basis heeft voor vervolg cursussen. Een introductie tot alle niet-metrische analysetechnieken die criminologen gebruiken wanneer zij afhankelijke variabelen hebben gemeten op ordinaal niveau of nominaal niveau en onafhankelijke variabelen op metrisch niveau valt buiten het bestek van deze inleidende cursus. Echter, onthoud dat er ordinale en nominale varianten van de regressieanalyse bestaan.<sup>14</sup>

---

<sup>14</sup> Een zeer vaak voorkomende analysetechniek wanneer de afhankelijke variabele van het nominale niveau is, is de binomiale logistische regressie. Van alle niet-metrische (multivariate en bivariate) analysetechnieken komt deze wellicht het meest voor in de criminologische wetenschappen. Binomiale logistische regressie analyse is geschikt voor een afhankelijke variabele die categorisch van aard is: er zijn maar twee categorieën. We behandelen deze techniek echter niet in het licht van deze syllabus.

## 7. Leerdoelen

Dit hoofdstuk beoogt diverse leerdoelen. Je wordt verwacht te weten onder welke omstandigheden je kiest voor een correlatie of voor een regressieanalyse. We verwachten dat je zelf (weliswaar met behulp van een rekenmachine) een correlatie- en regressieanalyse kan uitvoeren. Je dient alle coëfficiënten die betrekking hebben op zulke analyses goed te begrijpen: de ongestandaardiseerde en gestandaardiseerde richtingscoëfficiënt, het intercept, R-kwadraat, het begrip “model fit”. Maak daarom een inventaris van begrippen en tracht deze uit te leggen in je eigen woorden. Houd daarbij de statistische achtergrond in het oog.

### Samenvattend schema voor bivariate beschrijvende analysetechnieken

VERBANDEN TUSSEN 2 VARIABELEN: symmetrische associatiematen			
	Nominaal	Ordinaal	Metrisch
<i>verbanden tussen 2 variabelen</i>	<i>Chi<sup>2</sup>, Phi, Cramer's V</i>	<i>Spearman's rang-correlatie; Kendall's Tau-b, gamma</i>	<i>Correlatie-analyse</i>

### Dependente bivariate analysetechnieken

Onafhankelijke variabele	Afhankelijke variabele	ANALYSETECHNIEK	Parameters
<i>Interval/Ratio</i>	<i>Interval/Ratio</i>	Lineaire Regressieanalyse	Intercept Ongestandaardiseerde en gestandaardiseerde richtingscoëfficiënt Determinatiecoëfficiënt Regression Sum of Squares Residual Sum of Squares



## Hoofdstuk 9

### Inferentiële statistiek en variantieanalyse

#### **1. Waarom gebruiken we inferentiële statistiek?**

Wanneer we vragen stellen over daders van bepaalde types criminaliteit, over het effect van roken op longkanker,... dan hebben die vragen meestal betrekking op de hele populatie. Maar het is per definitie onmogelijk om alle leden van een bepaalde populatie te onderzoeken. Zoiets is bovendien vaak te tijdrovend en te kostelijk, zelfs al kan men de populatie identificeren. Er zijn situaties waarin we over populatiegegevens beschikken, maar voor een beperkt aantal kenmerken. We mogen tot slot ook niet vergeten dat we in het domein van de criminologie ook in die gevallen waarin men beschikt over populatiegegevens, voorzichtig moet zijn. Populatiedata over criminale feiten weerspiegelen een fractie van de totale criminaliteit. Onze data moeten steeds met de nodige voorzichtigheid worden benaderd.

We verkiezen steeds steekproeven die een zo getrouw mogelijke afspiegeling zijn van de populatie en die, in theorie, een excellente basis vormen om besluiten naar de ganse populatie door te trekken. Een steekproef is steeds een subgroep van de populatie die moet bestudeerd worden. Vooraleer we kenmerken in steekproeven gaan bestuderen, is het belangrijk dat we ons de vraag stellen hoe goed de steekproef is. Is de steekproef werkelijk een getrouwe en representatieve afspiegeling van de populatie. Speelt dit dan zo een rol? Natuurlijk! Want als je een steekproef neemt waarbij een belangrijk segment van de populatie niet werd bevraagd, of heel erg ondervertegenwoordigd is, wat heb je dan aan zulke resultaten? De kennis over de steekproeftrekking komt in andere opleidingsonderdelen aan bod. In dit handboek besteden we aandacht aan de mechanismen die we gebruiken om verbanden door te trekken naar een populatie op basis van een steekproef en gaan er vanuit dat onze steekproeven voldoen aan de kwaliteitsvooraarden. Als een steekproef niet voldoet aan de essentiële kwaliteitsvooraarden, zoals random-selectie en dergelijke meer, dan is het in feite gevaarlijk om gebruik te maken van de inferentiële statistiek.

Maar nu terug naar onze steekproeven en de inferentiële statistiek. In het onderzoeksproces verzamelen we steeds een beperkt aantal kenmerken voor elke onderzoekseenheid die tot de steekproef behoort: dit noemen we de ruwe data of ruwe gegevens (raw data in het engels). Deze ruwe data worden gebruikt om variabelen van te maken die bruikbaar zijn voor

criminologisch onderzoek. Criminologen proberen op basis van die partiële informatie afkomstig uit steekproeven uitspraken te doen over een gehele populatie. Om de resultaten op basis van deze steekproeven te kunnen veralgemenen naar de populatie, maakt men gebruik van de *inferentiële statistiek*, die gebaseerd is op de principes van het kansrekenen. Dankzij de inferentiële statistiek kunnen we uitspraken doen over een breder geheel aan eenheden dan diegenen die we bevraagd hebben. Dit doen we via het gebruik van *betrouwbaarheidsintervallen* en *significantietoetsen*. In dit deel staat *in de eerste plaats* het verwerven van inzicht in de basisprincipes van het infereren, i.e. het doen van uitspraken over een populatie op basis van een steekproef, centraal. *In de tweede plaats* staat de interpretatie van significantietoetsen en betrouwbaarheidsintervallen centraal.

## 2. De representativiteit van steekproeven

In dit deel van de cursus statistiek in de criminologie worden een aantal basisbeginselen van de inferentiële statistiek besproken. Het gaat hier slechts om een beperkte inleiding die ons toelaat om met enig inzicht te oordelen over het verwerken van gegevens en zonder te grote fouten een analyse uit te voeren. Deze inleiding volstaat echter niet om op eigen houtje een grondige en gefundeerde analyse van gegevens te kunnen uitvoeren.

Hiervoor hebben we reeds gesteld dat het doel van de *beschrijvende statistiek* bestaat in het op een overzichtelijke en synthetische wijze weergeven van alle gegevens die we voor een populatie of steekproef hebben verzameld. We ‘beschrijven’ deze gegevens van de steekproef of de populatie op verantwoorde wijze maar kunnen geen causale effecten (of oorzaak-gevolg) “bewijzen”. We kunnen wel aan de hand van beschrijvende statistiek mogelijke ideeën opdoen over de samenhang van kenmerken en hypotheses opstellen die dan via de *inductieve statistiek* moeten getoetst worden. Het doel van de inductieve statistiek bestaat uit het *veralgemenen* van de gegevens verzameld voor een steekproef naar de populatie waaruit ze getrokken werden. Inductie betekent dat we van het bijzondere (de steekproef) naar het algemene (de populatie) gaan. In onderzoek waar we stellingen uit theorieën toetsen vertrekken we vanuit algemene theorieën en willen die getoetst zien in een populatie. Dat is deductie: het afleiden van specifieke veronderstellingen uit algemene theorieën over criminaliteit. Als we dan de steekproefdata hebben verzameld en geanalyseerd, moeten we onze steekproefgegevens doortrekken naar de populatie en dat laatste is een inductief element. Dus: ook in deductief onderzoek zit altijd een stukje inductie.

Deze veralgemening van resultaten en het toetsen van hypothesen is echter niet eenvoudig. Er treedt vaak een aantal verstorende factoren op bij een onderzoek die een invloed uitoefenen. Zo kan men bv. door onprecieze metingen bepaalde onjuiste waarnemingen bekomen. Foutieve waarnemingen kunnen het gevolg zijn van slechte meetinstrumenten. Dat probleem bestaat ook in andere domeinen. Het is geen exclusief probleem van de sociale wetenschappen en de criminologie, echter, het is wel zo dat het in de sociale wetenschappen vaak iets moeilijker is om precieze metingen uit te voeren. Waarnemingen gedaan op een bepaald ogenblik in de tijd, kunnen later andere resultaten opleveren. We spreken van de ***meetfout***. Deze kan toevallig zijn of systematisch. Toevallige meetfouten zijn minder erg dan systematische meetfouten. Toevallige meetfouten heffen elkaar op, maar systematische meetfouten zorgen ervoor dat we de bal compleet mis slaan: we meten immers steeds onvolledig, omdat ons meetinstrument defect is. Vergelijk de situatie waarin je herhaaldelijk meet met een slecht afgestemd instrument. Bij herhaalde metingen heb je dezelfde uitkomst, maar deze is telkens fout. Zelfs indien we alle regels bij het trekken van steekproeven volgen, kunnen er vertekeningen optreden in de steekproef. Men spreekt van de ***steekproeffout***. Aangezien een steekproef slechts een deel van een populatie is, kunnen verschillen gemeten in de ene steekproef groter of kleiner zijn in de populatie. M.a.w. er zijn onzekerheden die inherent zijn aan het proces van het generaliseren naar velen op basis van weinigen. Statistische inferentie houdt zich met deze onzekerheden bezig door 2 functies te vervullen:

- **schatting:** het gebruik van steekproefkenmerken om ze naar de populatie te veralgemenen. bv. we voorspellen op basis van de meting in de steekproef die opleverde dat 49% vindt dat de regering haar werk goed doet, dat 49 % van de bevolking vindt dat de regering haar werk goed doet. Hoe precies is deze schatting? Steunend op de waarschijnlijkheidstheorie zullen we hier naar een antwoord zoeken.
- **hypothesen testen of toetsen:** kan men met voldoende zekerheid een bepaalde onderzoekshypothese aannemen of moet men ze op basis van de gegevens verwijderen?

In statistische inferentie zijn de vier vermelde begrippen (**populatie, steekproeven, schatting en toetsing**) cruciaal. We zullen ze daarom in dit deel één voor één toelichten en hier in eerste instantie blijven stilstaan bij het onderscheid populatie versus steekproef.

### **3. Steekproeven en populatie**

De relevante populatie voor een studie noemen we **de set (groep/reeks) van personen of objecten waarin een onderzoeker geïnteresseerd is**. Soms kunnen we gegevens verzamelen voor alle individuen uit een populatie, maar vaak wordt slechts een deel van de populatie onderzocht: dit deel noemen we de steekproef. Een eerste stap in onderzoek is zich de vraag te stellen wat de populatie is die men wenst te onderzoeken. Deze populatie kunnen we afleiden uit de onderzoeksfrage.

- *Voorbeeld 1. Heeft de detentieduur een effect op het zelfbeeld van gedetineerden? Hier bestaat de populatie uit alle gevangenenv.*
- *Voorbeeld 2. Wordt het zelfbeeld van personen die geïnterneerd zijn beïnvloed door processen van etikettering op basis van het type delict? Hier bestaat de populatie uit geïnterneerden.*
- *Voorbeeld 3. Wat zijn de belangrijkste demografische achtergrondkenmerken van slachtoffers van seksuele delinquentie? Hier bestaat de populatie uit alle slachtoffers van seksuele delinquentie.*
- *Voorbeeld 4. Genieten politiezones waar de “community policing” filosofie aanhangen wordt een hoger aanzien onder de bevolking dan politiezones waar dat niet het geval is. Hier is de populatie de bevolking. Als je dacht dat het de politiecommissariaten waren, dan heb je je laten misleiden. We zijn geïnteresseerd in de attitudes van burgers ten aanzien van politiezones. Community policing is een kenmerk van politiezones. De formulering heeft je wellicht misleid. Lees dus steeds goed de formulering.*
- *Voorbeeld 5. Worden werklozen sneller daders van criminaliteit dan werkenden? Hier bestaat de populatie uit werklozen en werkenden.*
- *Voorbeeld 6. Welke vormen van criminaliteit komen het meeste voor in België? Hier bestaat de populatie uit alle criminaliteit in België, dit wil zeggen alle feiten die een strafrechtelijke vervolging kunnen hebben. En dat zijn er heel wat! België heeft enorm veel bijzondere strafwetten. Sommigen stellen zich daarom de vraag of het systeem niet moet vereenvoudigd worden.*

Wanneer men een onderzoekspopulatie kent, moet men zich afvragen of het om een **afbakenbare** of een **hypothetische populatie** gaat.

### **Afbakenbare populaties**

In vele gevallen kan de populatie onder studie fysiek in kaart worden gebracht. De gegevens opgeslagen in computers of in een kaartenbakensysteem leveren ons lijsten. Zo worden alle mogelijke personen die mogen stemmen op lijsten bijgehouden, beschikt men over een aantal kenmerken van de bevolking via de bevolkingsregisters, houden gevangenissen lijsten bij van degenen die opgesloten worden, houden jeugdrechters dossiers bij van degenen die ze begeleiden en plaatsen, enz.

Bekijken we de eerder geformuleerde onderzoeks vragen dan blijken voorbeelden 1, 2, 4 en 5 over afbakenbare populaties te gaan.

### **Hypothetische populaties**

In het geval van het bestuderen van de invloed van roken op longkanker, de criminaliteitscijfers, cijfers over slachtofferschap,... is de populatie veel minder duidelijk afgebakend. (alle rokers, alle daders of slachtoffers, ....) Men noemt ze **hypothetische populaties** omdat we ze niet op lijsten terugvinden of kunnen identificeren. De populatie kan zelfs elementen en mensen inhouden die nu nog niet bestaan; bv. het aantal kinderen dat volgend jaar zal gemaakt worden. Filosofisch gezien kunnen volledige populaties waarover we gegevens hebben, bijvoorbeeld alle Belgen die in 2010 geregistreerd staan, ook beschouwd worden als een “steekproef van alle Belgische populaties uit alle voorbije jaren”. Toegegeven, het is misschien wat ver gezocht. Sommige criminologen zien hierin een reden om ook op populatiegegevens statistische significanties toe te passen. De meningen zijn hier evenwel over verdeeld. Wij menen alvast dat het geen kwaad kan om ook op populatiegegevens ook inferentiële statistiek toe te passen. Waarom? Wel, het argument dat vaak gebruikt wordt is dat een statistische significantietoets meer ziet dan het blote oog alleen. Wat daarmee bedoeld wordt, zal duidelijk worden wanneer we de variantieanalyse uiteenzetten.

## **4. Steekproeven en het principe van toeval**

Een steekproef moet zo goed mogelijk de relevante kenmerken van een populatie vertegenwoordigen; ze moet met andere woorden *representatief* zijn. In de praktijk zullen we steeds spreken over een *zeer goede afspiegeling* in plaats van een perfecte afspiegeling. Om een goede afspiegeling te krijgen zullen we de steekproef lukraak samenstellen (*'random sample' of 'toevalssteekproef'*). Door statistieken te berekenen of te rapporteren voor een steekproef in plaats van voor de gehele populatie, treedt er een fout op die **steekproeffout**

wordt genoemd. Deze steekproeffout wordt bepaald en onder controle gehouden door voldoende aandacht te besteden aan het steekproefkader, het steekproefontwerp en de implementatie ervan. Het steekproefkader geeft weer wie - administratief - deel uitmaakt van de te onderzoeken doelpopulatie, en bijgevolg kans heeft of moet hebben om in de steekproef opgenomen te worden. Zo beogen nationale enquêtes naar onveiligheidsbeleving en slachtofferschap in eerste aanleg representativiteit voor 'de bevolking van 15 jaar en ouder'.

Een steekproef is bovendien pas geslaagd indien zij op de relevante kenmerken waarin de criminoloog geïnteresseerd is, de populatie goed weerspiegelt. Men zegt dan dat de steekproef voor deze relevante kenmerken **representatief** is. 'Representativiteit' betekent dat de vermelde kenmerken met de juiste aantallen aanwezig zijn in de steekproef. Het heeft geen enkele zin, en is zelfs misleidend, om een steekproef zomaar - zonder de kenmerken te vermelden - 'representatief' te noemen en zo de indruk te wekken dat de steekproef voor alle kenmerken een goede weergave is; immers, dat is een steekproef vrijwel nooit. In de meer populariserende 'polls' of opiniepeilingen (zoals bijvoorbeeld in het peilen naar het kiesgedrag van mensen) wordt dit voortdurend gedaan om in de media de indruk te wekken dat wat verteld wordt, 'waar' is.

In de praktijk kan de representativiteit van een steekproef vaak maar worden aangetoond voor een beperkt aantal kenmerken, zoals sekse, leeftijd, regio. In criminologisch onderzoek is dit niet zelden des te problematisch aangezien crimineel gedrag of slachtofferschap in belangrijke mate verborgen is. Slachtofferenquêtes en zelfrapportagestudies pogen beiden om een bepaald "*dark figure*"-problematiek op te heffen. Deze studies worden in hoofdzaak uitgevoerd net omwille van het ontbreken van een externe en betrouwbare bron met betrekking tot de fenomenen van daderschap en slachtofferschap. Het spreekt voor zich dat de representativiteit van de steekproef, hoe relatief dit vaak is, hierbij uiteraard van cruciaal belang is.

## 5. De theorie van toevalssteekproeven

Toevalssteekproeven zijn **aselecte** steekproeven. Het zijn steekproeven waarin elke elementaire eenheid uit de empirische populatie een **bekende** (lees: **berekenbare**) **kans** heeft om in de steekproef opgenomen te worden. Wanneer men het toeval laat spelen, kan de kans berekend worden aan de hand van de spelregels van de kansrekening, die reeds in deze syllabus aan bod kwamen. Het basismodel van de toevalssteekproef is een **enkelvoudige aselecte steekproef** of 'simple random sample'. In de steekproeventheorie maakt men

doorgaans een onderscheid tussen **steekproefgrootheden** en **populatieparameters**. Statistieken afkomstig uit steekproeven zoals het gemiddelde, de standaardafwijking, een richtingscoëfficiënt,... zijn steekproefgrootheden. Populatieparameters verwijzen naar het gemiddelde, de standaardafwijking, een richtingscoëfficiënt,... in de populatie. Die waarde is per definitie niet gekend. We vertrekken immers van de steekproef om iets te zeggen over de populatie.

Bij de louter beschrijvende statistiek wordt er geen onderscheid gemaakt tussen steekproef en populatie. Bij **inferentiële statistiek** is het verschil tussen steekproef en populatie essentieel. De eigenschappen van de steekproef zijn **bekend** (*gemiddelde, minimum, maximum, spreiding, enz.*). De steekproef is op zich niet interessant. We gebruiken hem om iets te weten te komen over de populatie. De eigenschappen van de populatie zijn **onbekend** (*gemiddelde, minimum, maximum, spreiding enz.*). We trekken een steekproef om informatie te krijgen over een onbekende populatie.

In de inferentiële statistiek doen we uitspraken over de populatie op basis van wat we vinden in een **aselecte steekproef** uit die populatie. Anders gezegd: we gebruiken steekproefkenmerken om iets te zeggen over populatiekenmerken.  
De onbekende kengetallen van de populatie geven we aan met Griekse letters, bijvoorbeeld  $\mu$  (mu),  $\sigma$  (sigma) en  $\pi$  (pi). Zulke kengetallen worden de parameters van de populatie genoemd.

	<b>In de steekproef: steekproefkenmerk</b>	<b>In de populatie: parameter</b>
gemiddelde	$\bar{x}$	$\mu$
variantie	$s^2$	$\sigma^2$
standaardafwijking	$s$	$\sigma$
fractie	$p$	$\pi$
omvang	$n$	$N$

Zoals we daarnet al even kort aangehaald hebben, zijn er **twee hoofdactiviteiten** binnen de inferentiële statistiek:

- **Schatten:** we berekenen een steekproefkenmerk (bijvoorbeeld het aantal verschillende delicten waarvan men slachtoffer wordt) en gebruiken de waarde van dat kenmerk (bijvoorbeeld gemiddeld 2 delicten) om een uitspraak te doen over een populatiekenmerk (het gemiddeld aantal delicten waar men in de bevolking slachtoffer van wordt). Zo'n uitspraak is altijd gebaseerd op kansrekening. We schatten de parameter ‘gemiddelde leeftijd’ (symbool  $\mu$ ) met het steekproefkenmerk ‘gemiddelde leeftijd’ (symbool  $x$ ).

We onderscheiden de begrippen **puntschatting** en **intervalschatting**. Onder puntschatting verstaan we de schatting van een kenmerk in de populatie op basis van steekproefgegevens. De schatting van het gemiddeld aantal delicten waarvan iemand slachtoffer wordt, is een puntschatting. Onder intervalschatting verstaan we de marges waarbinnen we met een zekere graad van onzekerheid een puntschatting inschatten. We spreken dan van **betrouwbaarheidsintervallen**.

- **Toetsen:** we veronderstellen dat er een verband bestaat tussen leefstijl en beroving op straat van geweld. Wie ’s avonds vaker uitgaat, heeft meer kans om beroofd te worden op straat.

( → **hypothese** = er is een positieve correlatie tussen het aantal avonden in de week dat men zich in het nachtleven stort en het aantal keer dat men slachtoffer wordt van een beroving). We vinden een positieve correlatie in onze steekproef, maar dat betekent per definitie niet dat dit ook zo is in de populatie. Onze uitspraak is gebaseerd op **kansrekenen**. Kansrekening vormt zo de basis van de inferentiële statistiek. Als we op grond van een steekproefresultaat iets zeggen over een onbekende populatie is dat altijd een onzekere uitspraak. In dit deel van de syllabus leren we met deze onzekerheid om te gaan.

We kunnen de theorie die achter het schatten van de kans dat een steekproefgrootheid een goede weerspiegeling vormt van de onbekende populatieparameter (bijvoorbeeld de proportie jongens in een steekproef) duidelijk maken aan de hand van een **experiment**, waarbij we in omgekeerde volgorde te werk gaan. Dit kunnen we doen door uit een zekere populatie - stel alle criminologiestudenten die in de eerste bachelor aan de Ugent studeren - meerdere kleinere steekproeven van eenzelfde omvang - stel 100 eenheden - te trekken met bekende parameter, zoals het geslacht van de studenten. Het gaat hier om een werkelijke populatie. Via de

studentenadministratie weten we dat de proportie mannelijke studenten 40% bedraagt. De verkregen steekproefuitkomsten (de proportie jongens in elke afzonderlijke steekproef) zijn telkens **frequentieverdelingen**. De toevalsvariabele of steekproefuitkomst waarin we geïnteresseerd zijn, is de waarde van de proportie jongens in verschillende steekproeven van 100 eenheden.

*Indien we dergelijke oefening repetitief zouden uitvoeren, zou blijken dat naarmate het aantal toevalsteekproeven van dezelfde omvang (in ons voorbeeld is  $n = 100$ ) groter wordt, de concentratie van de uitkomsten rond de werkelijke waarde toeneemt.* De werkelijke waarde is de waarde in de populatie, die we kennen via de studentenadministratie. Het aantal steekproeven met uitkomsten aan de uitersten van de steekproevenverdeling neemt, relatief gesproken, af. Merk ook op dat de uitkomst van sommige particuliere steekproeven soms een heel eind verwijderd kan zijn van de populatiewaarde (de proportie in de originele steekproef die dienst deed als populatie).

Wat betekent deze informatie nu concreet? De vraag die zich stelt is immers of we met de informatie over de steekproevenverdeling in het achterhoofd tot een *betrouwbare schatting* kunnen komen van de - onder normale omstandigheden onbekende - populatiewaarde; kunnen we met andere woorden op basis van de steekproefverdeling van een singuliere toevalssteekproef een betrouwbare schatting maken van de prevalentie van een zeldzame ziekte, of van het herhaald slachtofferschap, of het aantal harde kernjongeren? Op basis van de kennis omtrent steekproevenverdelingen weten we dat de kans om heel ver van de populatiewaarde te zitten niet zo bijzonder groot is indien het toevalskarakter van onze steekproef werd gerespecteerd.

## 6. Kenmerken van steekproevenverdelingen

In de inferentiële statistiek spreken we steeds over een aantal theoretische kansverdelingen waarin berekend wordt welke kans met een bepaalde waarde van een stochastische variabele verbonden is. Een (theoretische) steekproevenverdeling is altijd een theoretische kansverdeling die de functionele relatie toont tussen de mogelijke waarden van een bepaalde statistiek, gebaseerd op een steekproef van  $n$  eenheden, en de kans (dichtheid) verbonden met elke waarde, voor alle mogelijke steekproeven van identieke omvang  $n$  die uit een specifieke populatie getrokken worden. Steekproevenverdelingen hebben enkele belangrijke

eigenschappen waar onderzoekers gebruik van maken wanneer zij kwantitatief onderzoek doen op basis van steekproeven.

- **Ten eerste:** Indien de parameter die we willen schatten een gemiddelde is, dan weten we dat de verwachte waarde van het steekproevengemiddelde gelijk is aan het populatiegemiddelde. Dit geldt bij uitbreiding ook voor een percentage (of proportie) successen, aangezien dit eigenlijk een gemiddelde is van een variabele met code 1 (succes of kenmerk aanwezig) en code 0 (geen succes of kenmerk afwezig).
- **Ten tweede:** De variantie  $\sigma_s^2$  van de steekproevenverdeling van gemiddelden van onafhankelijke steekproeven met omvang  $n$  is steeds gelijk aan de populatievariantie  $\sigma^2$  gedeeld door de steekproefomvang  $n$  (dus  $\sigma_s^2 = \sigma^2 / n$ ). De vierkantswortel uit de variantie van de steekproevenverdeling - m.a.w. de standaardafwijking van het gemiddelde van de steekproevenverdeling -, wordt de **standaardfout** (of SE, 'Standard Error') genoemd.

$$\text{Dus: } \text{SE} = \sigma_s = \sigma / \sqrt{n}$$

- **Ten derde:** naarmate de steekproef groter wordt, neemt de kans toe dat het steekproefgemiddelde dichter bij het populatiegemiddelde komt. Dit is de wet van de grote getallen.
- **Ten vierde:** zowel het populatiegemiddelde als de populatievariantie kan worden geschat door het steekproefgemiddelde en de steekproefvariantie indien de steekproefomvang voldoende groot is.
- **Ten vijfde:** het gemiddelde en de standaardfout van de steekproevenverdeling kunnen bijgevolg ook geschat worden op basis van de parameters van een particuliere steekproef van voldoende omvang.

Deze eigenschappen stellen ons in staat om een schatting te maken van een onbekende populatieparameter en de betrouwbaarheid ervan volgens informatie over de parameters in de steekproef. Centraal hierbij staat echter dat het steeds om een schatting gaat met, inherent daaraan verbonden, een bepaalde mate van onzekerheid. Om echter gebruik te kunnen maken

van bovenstaand theorema zouden we moeten weten, zonder zelf grote aantallen steekproeven te trekken, hoe de theoretische kansverdeling van een bij ons probleem passende steekproevenverdeling er uitziet.

In de inferentiële statistiek zijn een aantal theoretische steekproevenverdelingen met hun eigenschappen bekend en gedocumenteerd. De meest bekende is de standaard normale verdeling, waaraan we reeds in een afzonderlijk deel aandacht hebben besteed. Men weet dankzij de theoretische statistiek ook onder welke omstandigheden deze delingen van toepassing zijn. Zo kunnen bijvoorbeeld veel steekproevenverdelingen gebaseerd op steekproeven met een voldoende grote omvang  $n$  worden benaderd met de normale verdeling, ook al is de verdeling van het bestudeerde kenmerk in de populatie niet normaal. De criminoloog maakt hiervan gebruik bij het beantwoorden van probleemstellingen.

## 7. Het gebruik van de normale verdeling in de inferentiële statistiek

De normale verdeling is de 'koning onder de kansverdelingen', niet omdat zoveel kenmerken in de sociale werkelijkheid zo'n verdeling hebben. Waarom dan wel? Als we een reeks grote steekproeven trekken uit een populatie, dan weten we dat de steekproefgemiddelen normaal verdeeld zijn. Deze eigenschap van de '*steekproefgemiddelenverdeling*' is heel belangrijk in de inferentiële statistiek. Deze verdeling is symmetrisch en éénoppig, gekenmerkt door 'klokvorm' en wordt volledig bepaald door de parameters  $\mu$  (**populatiegemiddelde**) en  $\sigma$  (**standaardafwijking**)

We hebben eerder gezien dat de normale verdeling piekt en symmetrisch verdeeld is rond het gemiddelde. Om de normale verdeling toepasselijk te maken voor een veelheid aan empirische delingen, maakt men gebruik van transformaties waarbij elke score wordt uitgedrukt als een genormeerde afstand tot het gemiddelde van de verdeling. Dit zijn de **z-scores** waarbij men; zoals voorheen uiteengezet, het verschil neemt van een waarde en het gemiddelde en dit deelt door de standaardafwijking.

## 8. De centrale limietstelling

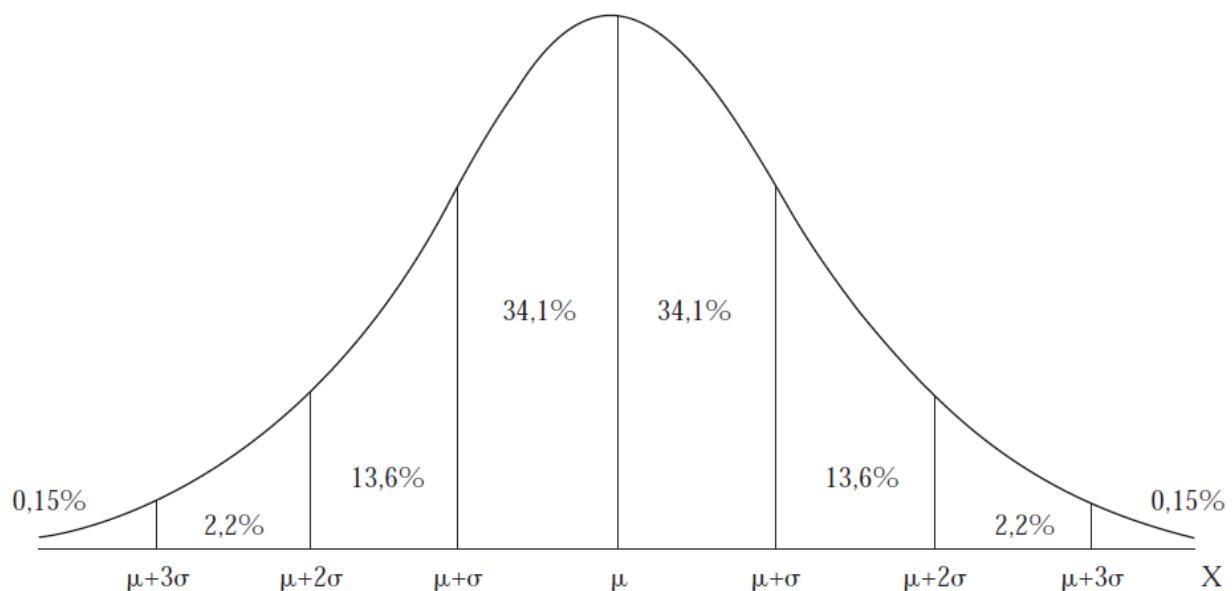
Voordat we de centrale limietstelling uitleggen herhalen we de drie bekende eigenschappen van de normale verdeling, met name de eerder besproken **68-95-99 regel**:

1. Ongeveer 68% van alle waarden valt binnen 1 standaardafwijking van het gemiddelde.
2. Ongeveer 95% van alle waarden valt binnen 2 standaardafwijkingen van het gemiddelde.
3. Ongeveer 99% van alle waarden valt binnen 3 standaardafwijkingen van het gemiddelde.

Het theorema van **de centrale limietstelling** is één van de belangrijkste begrippen uit de statistiek. Stel, je neemt heel veel steekproeven en van elke steekproef bepaal je het gemiddelde. Van al deze gemiddelden maak je een frequentieverdeling. Stel dat het gemiddelde van de hele populatie  $\mu$  is, dan vormen al die steekproefgemiddelden (bij voldoende grote steekproeven) een normale verdeling rond  $\mu$ . Dit is de centrale limietstelling.

Op basis hiervan kunnen we de kans berekenen dat alle steekproefgemiddelden in een bepaald interval rond  $\mu$  liggen. De Centrale Limietstelling is een versterking en precisering van de Wet van de Grote Getallen. Hoe groter dit aantal, hoe dichter de verdeling bij een normale verdeling ligt.

**Figuur: de normaalverdeling**



Wanneer we de normaalverdeling toepassen op de steekproevenverdeling dan is een z-score:

$$z_i = (x_i - \mu_s) / SE$$

Hierin stelt  $x_i$  elke mogelijke steekproefuitkomst voor.

Bij zo'n standaardnormaalverdeling (met gemiddelde 0 en standaardafwijking 1) ligt 99% van de oppervlakte onder de curve tussen  $-2,58z$  en  $+2,58z$  (eenheden standaardafwijking), 95 % tussen  $-1,96z$  en  $+1,96z$  en 90% tussen  $-1,64z$  en  $+1,64z$  eenheden standaardafwijking (de z-waarde). Met een kans van 5% op een vergissing ligt de werkelijke waarde met andere woorden tussen de waarden  $-1,96z$  en  $+1,96z$ . Met deze kans op vergissingen zijn sociale wetenschappers bereid te leven.

We kunnen steeds bepalen welke de kans is op een welbepaalde steekproefuitkomst indien we het populatiegemiddelde en de populatievariantie kennen. We kunnen dan immers de parameters van de steekproevenverdeling berekenen. In zo'n geval kunnen we ook met een vastgestelde waarschijnlijkheid de hypothese toetsen of een bepaalde steekproefuitkomst al dan niet uit die populatie afkomstig kan zijn. Indien we bovendien een voldoende grote toevalsteekproef getrokken hebben en we een populatieparameter of -statistiek (gemiddelde, proportie of percentage) willen schatten, kunnen we gebruik maken van de eigenschappen van de normaalverdeling.

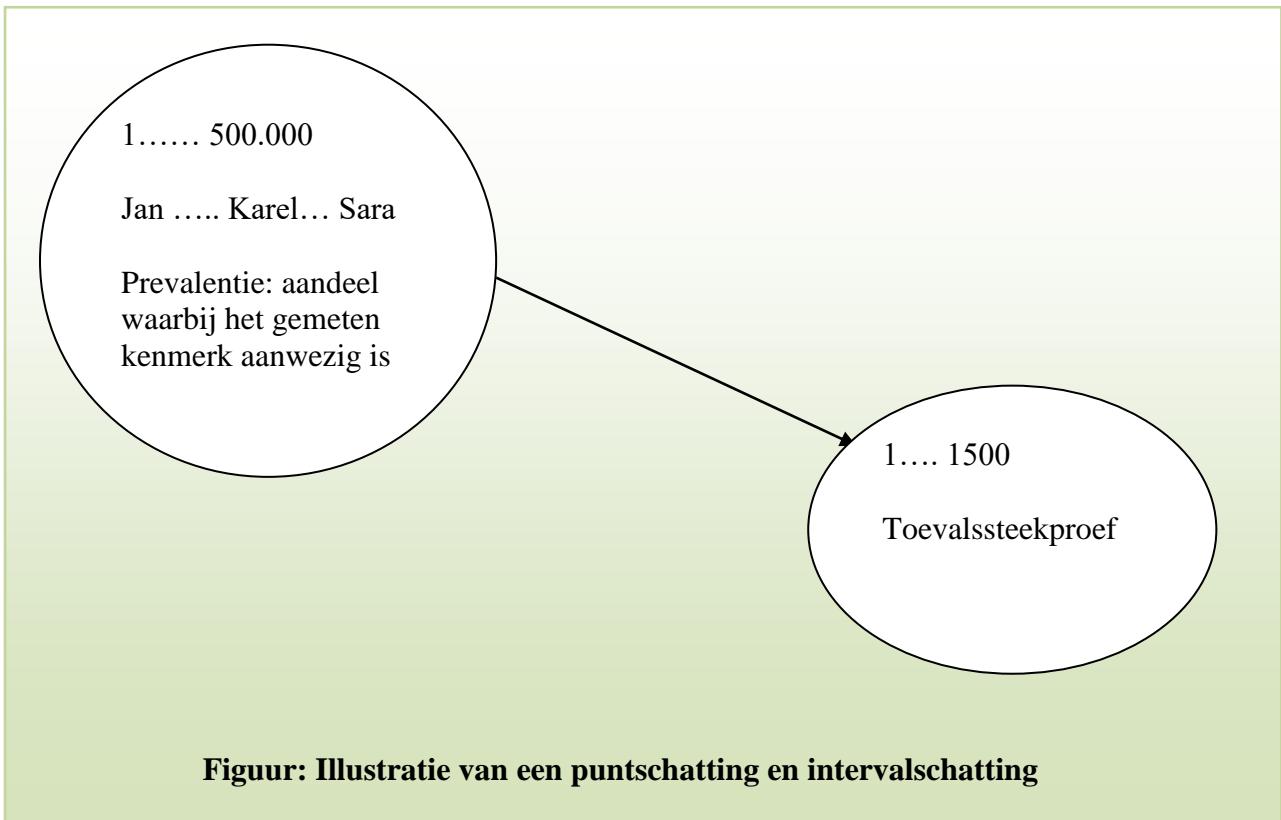
Deze procedure noemen we het *schatten en toetsen van parameters*. Met deze informatie is het mogelijk om, met een bepaalde waarschijnlijkheid en dus ook met een bepaalde kans op vergissing (bijvoorbeeld 5%), te bepalen binnen welk interval rond de steekproefuitkomst het populatiegemiddelde zal vallen. Zo'n interval wordt een **betrouwbaarheidsinterval of -gordel** genoemd. In feite zal elke veralgemening van een steekproefwaarde naar de populatie rekening moeten houden met een interval dat qua breedte varieert naargelang de steekproefomvang, de mate van spreiding in de populatie (standaardafwijking), de gewenste betrouwbaarheid en het type van toevalsteekproef.

## 9. Puntschatting en intervalschatting

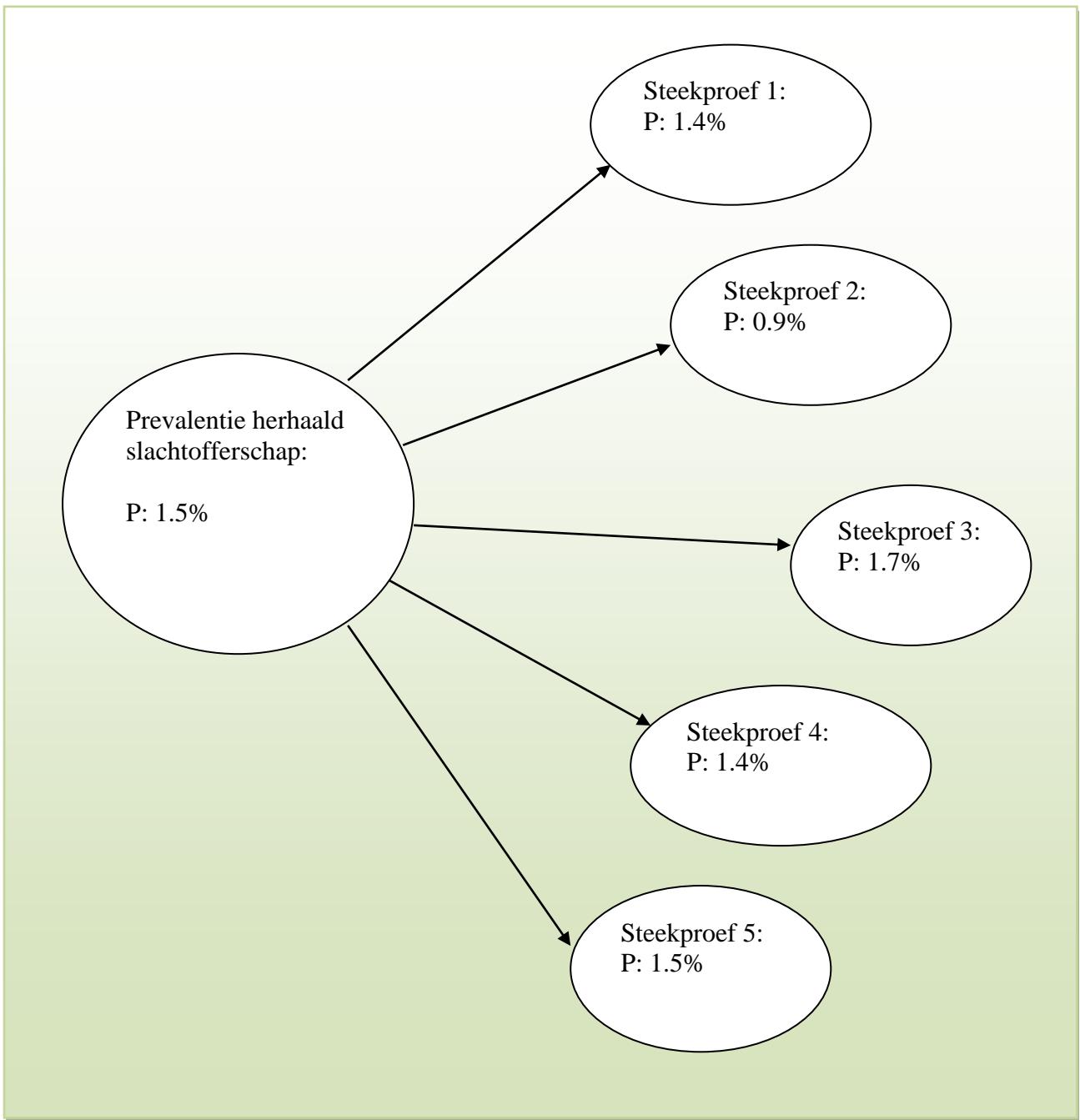
Uit het voorgaande bleek reeds dat statistische inferentie inhoudt dat men het resultaat van een steekproefonderzoek veralgemeent naar een volledige populatie. Belangrijke begrippen die men zeer goed dient te kennen zijn de begrippen *punt- en intervalschatting*. Zij vormen de basis van de statistische hypothesetoetsing en statistische inferentie, die verderop wordt uiteengezet. Met statistische inferentie kan men eigenschappen in populaties schatten met behulp van toevalsteekproeven. Veronderstel dat we geïnteresseerd zijn in herhaald slachtofferschap van geweldsdelicten in een populatie.

We gaan te werk in drie stappen.

- (1) We nemen een toevalssteekproef van 1500 individuen
- (2) We bevragen de individuen in de steekproef en noteren hoeveel individuen herhaald slachtoffer werden van een geweldsdelict
- (3) We schatten de prevalentie



Uit de toevalssteekproef blijkt dat 18/1500 het kenmerk bezitten. Dit is 1.2%. De prevalentie van het kenmerk ‘herhaald slachtofferschap’ wordt dus op 1.2% geschat. Dit is een puntschatting van de prevalentie met een zekere mate van onzekerheid. Om deze onzekerheid te illustreren, volgt hieronder een computersimulatie. We nemen vijf toevalssteekproeven waarbij we de prevalentie schatten. De populatie bestaat uit 500.000 individuen, waarvan 1.5% herhaald slachtoffer zijn geworden van een geweldsdelict.



**Figuur: Illustratie van een puntschatting en intervalschatting (vervolg)**

Een schatter is een grootheid die berekend wordt aan de hand van een bepaalde functie (algoritme) op basis van de informatie van de steekproef en wordt uitgedrukt in een formule. Een schatter is het resultaat van de schatting. We schatten steeds onbekende parameters in een populatie en dit op basis van informatie uit de steekproef.

Een *zuivere schatter* of *onvertekende schatter* (Engels: *unbiased estimate*) is dus een statistische grootheid waarvan de verwachtingswaarde samenvalt met de te schatten populatieparameter.

De geschatte prevalenties variëren tussen 0.9% en 1.7%. Hoe groot de variatie van de geschatte prevalenties tussen de verschillende steekproeven is, hangt samen met hoe groot de steekproeven zijn. Grottere steekproeven leiden tot een kleinere variatie in de prevalentieschattingen. We krijgen dus een zekerder resultaat als de steekproefgrootte toeneemt. Dit is een belangrijke statistische wetmatigheid.

In de praktijk van het onderzoek nemen we uiteraard genoegen met één steekproef. Op basis van deze steekproef zal een *intervalschatting* worden gemaakt. De intervalschatting geeft de (on)zekerheid van onze uitkomst weer. De schatting van het interval gebeurt aan de hand van **betrouwbaarheidsintervallen**. In het voorbeeld hebben we de prevalentie van het fenomeen herhaald slachtofferschap van een geweldsdelict geschat op 1.2%. Deze schatting geeft ons nog geen idee over de nauwkeurigheid ervan. Hoeveel vertrouwen kunnen we nu hebben in dit resultaat? De schatting van een betrouwbaarheidsinterval biedt ons meer precieze informatie. Een betrouwbaarheidsinterval is een schatting +/- een foutenmarge. Dit interval wordt berekend uit de steekproefdata volgens een methode die een bepaalde kans heeft een interval op te leveren waarin de populatiwaarde ligt. We baseren ons dus op het kansrekenen. Intervalschatting komt in het voorbeeld dus neer op het schatten van een interval dat ligt rond de geschatte proportie herhaalde slachtoffers in onze steekproef, en dat met een zekere graad van waarschijnlijkheid ook de onbekende populatieproportie van slachtoffers bevat.

De breedte van het interval geeft informatie over de precisie. De breedte van het interval kan worden berekend aan de hand van onze kennis over de steekproevenverdelingen van schattingen. Het komt er met andere woorden op neer dat, willen we die 1.2% veralgemenen naar de populatie, dit enkel kan in een uitspraak als volgt: ‘met 95% zekerheid kunnen we stellen dat de prevalentie van herhaald slachtofferschap van een geweldsdelict in de populatie ligt tussen de ...% en ...%’; het eerste percentage is dan ‘1.2% min het betrouwbaarheidsinterval’, het tweede percentage ‘1.2% plus het betrouwbaarheidsinterval’.

## **10. Het berekenen van een betrouwbaarheidsinterval rond een parameter**

**BI = schatting +/- foutenmarge**

**= interval berekend uit steekproefdata volgens methode die bepaalde kans heeft een interval op te leveren waarin de populatiewaarde ligt.**

Stel: we vinden een gemiddelde waarde  $\bar{x}$  in een steekproef. We kennen de steekproefgrootte, we kiezen een niveau van betrouwbaarheid (bijvoorbeeld 95%, dus hebben we 5% kans op een verkeerde waarde), en een foutenmarge (deze noemen we  $\alpha$ ) waarmee we bereid zijn te leven (in geval van 95% betrouwbaarheid is  $\alpha = 5\%$ ). Algemeen kunnen we hieruit berekenen tussen welke waarden de populatieparameter  $\mu$  zal liggen en kan het betrouwbaarheidsinterval als volgt weergegeven worden:

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

Daarin staat  $\bar{x}$  voor de steekproefuitkomst (bijvoorbeeld een gemiddelde, proportie of percentage).  $z_\alpha$  is een score van de standaard normaalverdeling die overeenkomt met de zelf gekozen kans op fout ( $\alpha$ ). Deze score wordt vermenigvuldigd met de standaardafwijking in de populatie gedeeld door de wortel van  $n$ . De standaardafwijking van het bestudeerde kenmerk in de populatie kennen we natuurlijk niet, maar we weten dat bij voldoende grote steekproeven de standaardafwijking in de steekproef wel kan gebruikt worden als schatter voor deze onbekende.

Een betrouwbaarheidsniveau CI (*Confidence interval*) wordt gedefinieerd als  $1-\alpha$  (waarbij geldt dat  $\alpha =$  de kans op een vergissing, bvb 5% of 10%).

Er zijn verschillende betrouwbaarheidsniveaus.

**CI= 90% (z-score: 1.645)**

**CI= 95% (z-score: 1.960)**

**CI= 99% (z-score: 2.576)**

Als we een uitspraak willen doen met een waarschijnlijkheid van 95% en een kans van 5% op vergissing, dan is  $z$  gelijk aan 1,960. Willen we de betrouwbaarheid van onze uitspraken

optrekken tot bijvoorbeeld 99 % - we permitteren ons dan slechts 1 % kans op een fout -, dan is  $z$  gelijk aan 2,576. Het bredere betrouwbaarheidsinterval compenseert in dit geval ons verlangen om met een grotere waarschijnlijkheid uitspraken te poneren.

### **Voorbeeld zelf uitrekenen betrouwbaarheidsintervallen:**

Uit een enquête afgenomen bij 3709 studenten blijkt, dat de gemiddelde score op het einde van het 6<sup>de</sup> middelbaar 72% bedraagt. De standaardafwijking in de populatie bedraagt 8.

- a) Geef het 95% en 99% betrouwbaarheidsinterval
- b) Bereken het 95% betrouwbaarheidsinterval voor dezelfde resultaten maar met  $n=10$ ,  $n=100$  en  $n = 1000$

*Oplossing:*

Gegeven:

$$x_{\text{gem}} = 72$$

$$\sigma = 8$$

$$N = 3709$$

a)

**95% betrouwbaarheid  $\rightarrow z = 1,96$**

$$72 - (1,96 * 8/\sqrt{3709}) < \mu < 72 + (1,96 * 8/\sqrt{3709})$$

$$71,74 < \mu < 72,26$$

**99% betrouwbaarheid  $\rightarrow z = 2,575$**

$$72 - (2,575 * 8/\sqrt{3709}) < \mu < 72 + (2,575 * 8/\sqrt{3709})$$

$$71,66 < \mu < 72,34$$

b)

**95% betrouwbaarheid**

**$\rightarrow z = 1,96$**

- Voor  $n=10$

$$72 - (1,96 * 8/\sqrt{10}) < \mu < 72 + (1,96 * 8/\sqrt{10})$$

$$72 - 4,9584 < \mu < 72 + 4,9584$$

$$67,04 < \mu < 76,96$$

- **Voor n= 100**

$$72 - (1,96 * 8/\sqrt{100}) < \mu < 72 + (1,96 * 8/\sqrt{100})$$

$$72 - 1,568 < \mu < 72 + 1,568$$

$$70,43 < \mu < 73,57$$

- **Voor n= 1000**

$$72 - (1,96 * 8/\sqrt{1000}) < \mu < 72 + (1,96 * 8/\sqrt{1000})$$

$$72 - 0,4959 < \mu < 72 + 0,4959$$

$$71,50 < \mu < 72,50$$

Ook statistische verwerkingspakketten berekenen de betrouwbaarheidsintervallen en geven deze weer. Van de studenten wordt verwacht dat ze de basisprincipes die er achter schuilgaan kennen en dat ze deze intervallen naar waarde leren schatten. Studenten dienen deze ook zelf te kunnen berekenen. Betrouwbaarheidsintervallen zijn een manier om met onnauwkeurigheid om te gaan. De grootte van het interval wordt bepaald door de standaardafwijking van het kenmerk in de populatie en ook door de grootte van de steekproef. Bij kleine steekproeven zullen de intervallen dus steeds groter zijn dan bij grote steekproeven.

## **11. Statistische hypothesetoetsing**

Nu we de theorie over de steekproevenverdeling en de schatting van populatieparameters op basis van steekproefuitkomsten uiteengezet hebben, gaan we verder in op de principes en de praktijk van het statistisch toetsen. Het komt er voor criminologen die in de praktijk onderzoek doen op basis van steekproeven, immers op neer goed te begrijpen wat een statistische toets ons kan leren. Met statistische hypothesetoetsing kan men testen hoe aannemelijk een bepaalde uitspraak over de populatie, op grond van de steekproef in kwestie, werkelijk is. Is een nieuwe behandeling van jeugdige delinquenten effectiever dan een oude behandelingsmanier? Is de proportie die herhaald slachtoffer wordt van een geweldsdelict werkelijk lager dan 1.4% van de bevolking?

Een *significantietoets* is een procedure om gegevens (zoals) uitkomsten van een steekproef te vergelijken met een vooropgestelde hypothese, die we de *nulhypothese* gaan noemen. Een hypothese is een bewering over parameters in een populatie. We kunnen bijvoorbeeld de hypothese stellen dat een parameter in de populatie niet verschilt van nul. In onze steekproef verschilt deze echter wel van nul. Kunnen we daarom aannemen dat dit in de populatie ook zo zal zijn? De uitkomst van een significantietoets leert ons meer. De uitkomst van een

significantietoets wordt uitgedrukt in termen van een kans die aangeeft hoe goed data en hypothese met elkaar overeenkomen.

Vragen kunnen zijn: is een effect aanwezig van variabele X op Y? Is er een associatie tussen het roken en krijgen van kanker? Is er een relatie tussen leeftijd en onveiligheidsbeleving? De hypothese die stelt dat het effect of een associatie niet bestaat noemen we de *nulhypothese* ( $H_0$ ). De *alternatieve hypothese* ( $H_a$ ) wordt door de onderzoeker geformuleerd. Significantietoetsen worden uitgevoerd om de sterkte van het bewijs tegen de nulhypothese vast te stellen.

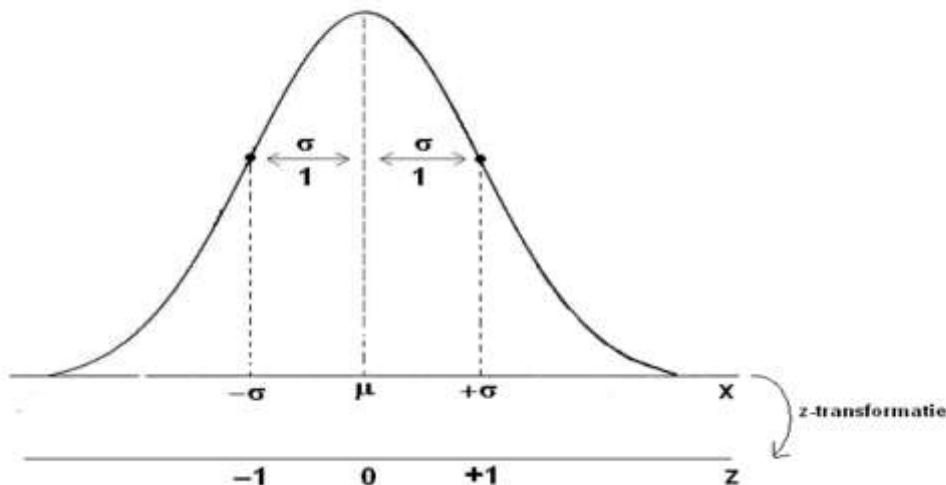
Hoe verder de waargenomen uitkomst verschilt van de nulhypothese, hoe onwaarschijnlijker dat  $H_0$  waar is, hoe sterker de indicatie voor  $H_a$ . De significantietoets meet de kans op het krijgen van een uitkomst die even extreem is, of nog extremer dan de waargenomen uitkomst. Dit noemen we de **overschrijdingskans** (p) van de toets.

Als  $p <$  is dan 0.05 betekent dit dat we, gesteld dat we een oneindig aantal steekproeven zouden trekken van dezelfde grootte, we slechts in 5 op 100 gevallen de gevonden steekproefuitkomst zouden uitkomen, terwijl deze in realiteit een waarde 0 heeft, of een andere vooropgestelde waarde.

Laat ons een voorbeeld geven uit de praktijk van het onderzoek. De onderzoeksvraag is of fraudeurs een significant hoger IQ hebben dan de gemiddelde mens, dit wil zeggen dat hun IQ hoger is dan op grond van louter toeval mag verwacht worden. We zetten de IQ scores van fraudeurs om in z-scores waardoor het gemiddelde 0 wordt en de standaardafwijking 1. We kennen de eigenschappen van de normale verdeling (oppervlakte: 1 of 100%, éénoppig en perfect symmetrisch) en kunnen opzoeken hoeveel procent van de fraudeurs een bepaald IQ heeft. Stel dat we uitgerekend hadden dat in onze steekproef de fraudeurs een gestandaardiseerd gemiddelde van 1.99 hebben. Boven de 1.99 bevindt zich slechts 2.33% van de z-scores van de normaalverdeling. De kans om dit resultaat te vinden, indien fraudeurs echt een gemiddeld IQ hebben is dus dermate klein, dat we de nulhypothese gaan verwerpen. Dit betekent dat we sowieso kunnen aannemen dat de fraudeurs uit onze steekproef significant afwijken van de gemiddelde Belg. Onze alternatieve hypothese (fraudeurs hebben een hoger IQ dan de gemiddelde Belg) wordt niet ontkracht. In statistisch jargon: onder de aanname dat er geen verschil is tussen fraudeurs en gemiddelde Belgen, is de kans op dit steekproefresultaat kleiner dan 5%. We verwerpen dus de aanname dat fraudeurs een

gemiddeld IQ hebben en gaan ervan uit dat fraudeurs een significant hoger IQ hebben dan de gemiddelde Belg.

$\underline{x} \sim N(\mu, \sigma)$  en  $N(0,1)$  in één figuur



Statistische hypothesetoetsing kan in zekere zin vergeleken worden met de uitspraak die een strafrechter doet over een beklaagde. De nulhypothese is de hypothese die we willen verwerpen. De nulhypothese in het geval van een rechter is de uitspraak dat de beklaagde onschuldig is. De rechter moet schuld als bewezen achten en dus moet de nulhypothese van onschuld ontkracht worden. Parallel kan men stellen dat de nulhypothese veronderstelt dat een bepaalde uitkomst of een steekproefgrootheid uit onze steekproef, eigenlijk nul is. De alternatieve hypothese is dat de beklaagde wel schuldig is; de alternatieve hypothese is dat een steekproefgrootheid verschilt van nul.

#### Schema: Het oordeel van de strafrechter

	Vrijlaten van de beklaagde	Veroordelen van de beklaagde
<b>De beklaagde is onschuldig (Nulhypothese)</b>	De rechter laat een onschuldige beklaagde vrij. <b>Correct besluit</b>	De rechter veroordeelt een onschuldige beklaagde. <b>Fout besluit en type-I fout.</b>
<b>De beklaagde is schuldig (Alternatieve hypothese)</b>	De rechter laat een schuldige beklaagde vrij. <b>Fout besluit en type-II fout</b>	De rechter veroordeelt een schuldige beklaagde. <b>Correct besluit</b>

Een *type-I fout* maken houdt in dat men een correcte nulhypothese verwerpt. Een *type-II fout* maken betekent dat men een foute nulhypothese aanvaardt. De testvariabele is de variabele

waarvan een waarde berekend wordt op basis van een observatie uit de steekproef. Het kan ook gaan om een samenhang tussen kenmerken (zoals een correlatiecoëfficiënt of een regressiecoëfficiënt). Bij hypothesetoetsing hoort een **p-waarde** ( $p = \text{probabiliteit}$ ). De p-waarde kan worden gezien als een uitdrukking van de waarschijnlijkheid of het ‘waarheidsgehalte’ van een nulhypothese. Zo is bijvoorbeeld een hoge p-waarde een indicatie dat de in de steekproef aangetroffen samenhang tussen twee kenmerken weinig waarschijnlijk is en de nulhypothese of de veronderstelling dat er geen samenhang bestaat tussen beide kenmerken in de populatie, meer ondersteuning geniet.

Of nog algemener kunnen we de mogelijke beslissingen uit steekproefinformatie als volgt voorstellen:

### Beslissing uit een steekproef

	Aanvaarden van de nulhypothese	Verwerpen van de nulhypothese
<b>Toestand A: De nulhypothese is juist</b>	<b>Correct besluit</b> $1-\alpha$	<b>Verwerpingsfout. Type-I fout.</b> <b>Vals alarm: de brandmelder gaat af, maar er is vals alarm</b>
<b>Toestand B: De nulhypothese is fout</b>	<b>Doorlatingsfout en type-II fout</b> <b>Niet ontdekken dat er iets bijzonders aan de hand is</b> <b>De brandmelder gaat niet af, terwijl er wel brand is</b>	<b>Correct besluit</b> $1-\beta^{15}$

Een statistische test (gebaseerd op de verzamelde gegevens) start met het stellen van een nulhypothese (' $H_0$ '). Men gaat na of deze nulhypothese waar of vals is, en of men de nulhypothese met andere woorden al dan niet dient te verwerpen. Terwijl de nulhypothese stelt dat een steekproefgrootheid niet significant verschilt van nul, is de **p-waarde** een indicatie van de waarschijnlijkheid van deze nulhypothese. Hoe weten we nu of een

---

<sup>15</sup> Dit gebied,  $1 - \beta$  noemt men ook de power van de test. Het is het geheel van waarden waaronder een valse nulhypothese correct wordt verworpen en een juiste nulhypothese wordt behouden. De power van een test is de kans op het detecteren van een effect wanneer er in werkelijkheid één is, en het besluiten dat er niets aan de hand is, wanneer er werkelijk niets aan de hand is.

In de praktijk neemt men aan dat toetsen met een power die groter is dan .80 (of  $\beta \leq .2$ ), krachtige toetsen zijn. In deze inleidende syllabus gaan we niet verder in op deze details.

nulhypothese waar of vals is? Indien de p-waarde kleiner is dan 0.05 (bij een  $\alpha$  van 0.05 of een lagere  $\alpha$  van 0.01, afhankelijk van de beslissing van de onderzoeker), dan is de kans dat de gestelde nulhypothese waar is, kleiner dan 5%. Dit is niet veel, en de meeste sociale wetenschappers zijn bereid deze grens als ondergrens te beschouwen: is de p-waarde met andere woorden kleiner dan 0.05, dan verwerpt men de nulhypothese. Softwarepakketten voor statistische analyse vermelden de exacte significantie in de output. Het is belangrijk te weten dat men die overschrijdingskans blijvend dient te zien als een kans en niet als echt bewijs.

In de samenvattende tabel hieronder geven we 4 voorbeelden van correlaties tussen X en Y uit een steekproef

### Samenvattende tabel

	Aanvaarden van de nulhypothese	Verwerpen van de nulhypothese
<b>H0 = waar</b>	$H_0 = 0$ $r = 0.09$ $p: > .05$ aanvaarden $H_0$ = JUISTE BESLISSING	<b><math>H_0 = 0</math></b> <b><math>r = 0.20</math></b> <b><math>p: &lt; .05</math></b> <b>verwerpen <math>H_0</math></b> <b>= TYPE 1 FOUT</b>
<b>H0 = niet waar</b>	<b><math>H_0 \neq 0</math></b> <b><math>r = 0.10</math></b> <b><math>p: &gt; .05</math></b> <b>aanvaarden <math>H_0</math></b> <b>= TYPE 2 FOUT</b>	$H_0 \neq 0$ $r = 0.32$ $p: < .05$ verwerpen $H_0$ = JUISTE BESLISSING

Samengevat:

**Stappen bij een significantietoets :**

**Formuleer de nulhypothese  $H_0$  en de alternatieve hypothese  $H_a$**

**Specificeer het significantieniveau  $\alpha$  (bijvoorbeeld een foutenmarge van 5% Type I fout)**

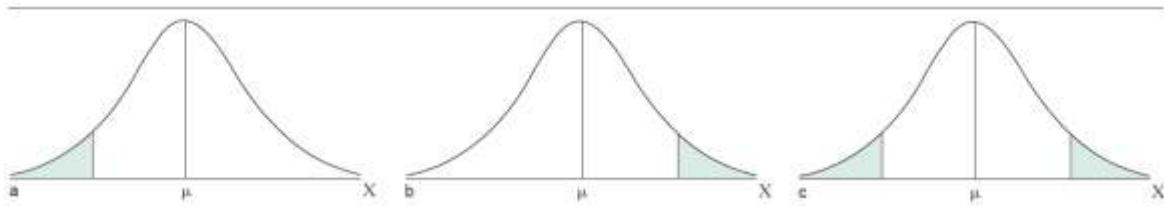
**Bereken de steekproefgrootte ( bereken de correlatie, het gemiddelde, ...)**

**Bepaal de bijhorende p-waarde of de overschrijdingskans. Is de p-waarde kleiner of gelijk aan  $\alpha$ , dan is het toetsresultaat statistisch significant op niveau  $\alpha$**

## 12. Eenzijdig of tweezijdig toetsen van een nulhypothese?

De formulering van de alternatieve hypothese bepaalt of we **eenzijdig** of **tweezijdig** toetsen. Een tweezijdige toets houdt in dat we stellen via de alternatieve hypothese ( $H_1$ ) dat de populatieparameter significant verschilt van nul, zonder een uitspraak te doen over de richting (positief verschillend van nul of negatief verschillend van nul). We spreken dan eigenlijk beter van toetsen met tweezijdig alternatief. Als de formulering daarentegen is, dat de populatieparameter positief verschilt van nul, dan wordt eenzijdig getoetst (beter: getoetst met eenzijdig alternatief). Het alternatief is rechtseenzijdig als de waarde van de populatieparameter een positieve waarde heeft en linkseenzijdig als de waarde van de populatieparameter een negatieve waarde heeft. Hoe we de alternatieve hypothese bij het toetsen formuleren en of we dus eenzijdig of tweezijdig toetsen hangt af van de onderzoeks vraag.

**Figuur: eenzijdig toetsen (links of rechts) en tweezijdig toetsen**



De p-waarde kan ook worden gezien als de kans dat een bepaald steekproefgemiddelde (of de correlatie) geheel bij toeval tot stand kwam. Hoe kleiner deze p-waarde, des te minder er sprake kan zijn van een toevallig resultaat. Het is echter van belang zich niet blind te staren op de significantie van resultaten alleen. De statistische significantie van een steekproefresultaat wordt sterk bepaald door de grootte van de steekproef. Men dient daarom ook naar de grootte van het verband of effect te kijken. Ook zeer geringe en inhoudelijk weinig relevante correlaties kunnen bij een voldoende grote steekproef significant zijn. Verder kan een nulhypothese ook altijd foutief verworpen worden. In het voorbeeld van het oordeel van de rechter wordt dan een onschuldige beklaagde veroordeeld. Men spreekt van een type-I fout. Behoudt men de nulhypothese ten onrechte, dan is er sprake van een type-II fout.

Wat is nu de “**power**” van een test? Dit is de kans dat de statistische test leidt tot een correcte verwerving van een valse nulhypothese. De statistische power van een test is dus de mate waarin een test er in slaagt een echt effect te detecteren als het effect ook echt bestaat, rekening houdende met zowel type-I als type II fouten. Als de statistische analyse van de

power van een test een waarde van 0.80 of beter oplevert, dan wordt vaak aangenomen dat er voldoende power is. De waarde van de power varieert van 0 tot 1.

### 13. Andere belangrijke verdelingen

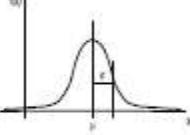
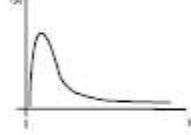
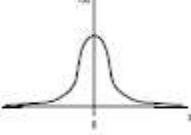
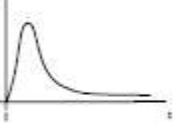
Niet alle kenmerken volgen een normale verdeling. Enkele andere belangrijke verdelingen, die eerder door wiskundige statistici werden beschreven, zijn belangrijk in het kader van de statistische inferentie. Deze andere verdelingen doen beroep op het begrip *vrijheidsgraden*, dat eerder al eens ter sprake kwam bij het uiteenzetten van de wijze waarop de steekproefstandaardafwijking wordt berekend. Wat onthouden dient te worden is dat van al deze andere verdelingen ook de oppervlakte onder de curve bekend is. We geven een overzicht van andere belangrijke statistische verdelingen. Sommige statistische significantietoetsen zijn gebaseerd op deze verdelingen.

**De chi-kwadraat ( $\chi^2$ )-verdeling** is een verdeling van het kwadraat van normaal verdeelde variabelen. Deze wordt onder andere gebruikt bij de verdeling van steekproefvarianties. De chi-kwadraat-verdeling is asymmetrisch en afhankelijk van de graden van vrijheid. Bij de steekproefvariantie zijn deze graden van vrijheid (degrees of freedom, afgekort Df) gelijk aan de steekproefgrootte (n) minus 1. De chi-kwadraat verdeling is in criminologisch onderzoek vooral belangrijk bij de analyse van kruistabellen om te weten of een percentageverschil statistisch significant is. De chi-kwadraat waarde die berekend wordt bij de analyse van kruistabellen volgt een chi-kwadraat verdeling.

De **t-verdeling** (Student's t) is een verdeling die uiterlijk heel erg lijkt op de normale verdeling. De frequentieverdeling (door wiskundigen *kansdichtheidsfunctie* genoemd) heeft dezelfde symmetrische klokvorm als die van de normale verdeling. De breedte van de klokvorm varieert (in tegenstelling tot bij de normale verdeling) met het aantal graden van vrijheid oftewel de "*degrees of freedom*" (**DF = de steekproefgrootte (n) minus 1**). De Student t-verdeling wordt gebruikt in de inferentiële statistiek waarbij we op basis van informatie uit één steekproef veralgemeeningen willen bekomen naar de populatie toe. In de t-tabel staan zowel voor eenzijdige situaties met kans  $\alpha$  in de staart (het oppervlak onder de curve rechts van t), als voor tweezijdige situaties met twee keer de kans  $\frac{1}{2} \alpha$  in de staarten (de som van de oppervlakten links van  $-t$  en rechts van t) de t-waarden. De tabel geeft dus zowel voor  $P(T < t) = (1 - \alpha)$  als  $P(-t < T < t) = (1 - \alpha)$  de t-waarden. De t-verdeling wordt altijd gebruikt bij de toets van richtingscoëfficiënten.

**De F-verdeling** (genoemd naar de statisticus Fisher) is een quotiënt van twee chi-kwadraat verdeelde grootheden. Deze wordt onder andere gebruikt bij het quotiënt van twee steekproef varianties uit twee normaal verdeelde populaties. De F-verdeling is net als de chi-kwadraat verdeling asymmetrisch. Bovendien is de vorm afhankelijk van graden van vrijheid ( $df = n-1$ ) in de teller en in de noemer van het quotiënt van variabelen dat samen F-verdeeld is. De F-verdeling wordt onder andere gebruikt in situaties waarin men wil weten of een determinatiecoëfficiënt die gevonden wordt in een steekproef statistisch significant verschilt van nul. De F-verdeling wordt gebruikt bij de toets van een **determinatiecoëfficiënt**.

### Figuur: enkele statistische delingen

	NV	$\chi^2$	Student	Fisher
Definitie	$Z \sim N(\mu, \sigma)$	$G_i \sim N(0, 1)$ $\chi^2 = G_1^2 + G_2^2 + \dots + G_n^2$	$G \sim N(0,1), Z \sim \chi^2_n$ $T = \frac{G}{\sqrt{Z/n}}$	$Z_1 \sim \chi^2_{n_1}, Z_2 \sim \chi^2_{n_2}$ $F = \frac{Z_1/n_1}{Z_2/n_2}$
Grafiek				
Parameters	$\mu, \sigma$	$n$ (vrijheidsgraden)	$n$ (vrijheidsgraden)	$n_1, n_2$
Gemiddelde	$\mu$	$n$	0	$\frac{n_2}{n_2 - 2}$
Eigenschap	symmetrisch ( $\mu$ )	$> 0$	symmetrisch (0)	$> 0$

De *binomiale verdeling* of **Bernoullie-verdeling** is genaamd naar wat het Bernoullie experiment genoemd wordt. Een Bernoullie experiment is een experiment met slechts twee mogelijke uitkomsten, die we “succes” ( S ) en “mislukking” ( M ) noemen. De binomiale verdeling wordt gebruikt bij de inferentie van *odds* en *odds-ratio’s*. Dit valt echter buiten het bestek van deze inleidende cursus.

In bijlage bij dit handboek vindt de lezer de tabellen of tafels waarbij de oppervlakte onder de curves vermeld staan. In de praktijk van het criminologisch onderzoek zullen we zien dat het

volstaat om de kritische waarden die we per verdeling vinden te vergelijken met de gevonden waarden in onze steekproef.

## 14. De variantieanalyse als toets voor verschillen tussen groepen inzake metrische kenmerken

Variantieanalyse (Engels: *analysis of variance*, ANOVA) is een toets voor de relatie tussen een **nominale** en een **metrische** variabele (bijvoorbeeld de relatie tussen verstedelijking en criminaliteit op basis van een gemeentetypologie). De berekeningswijze is gebaseerd op de **variaties van de steekproeven**. We geven een historisch voorbeeld: we doen een onderzoek naar criminaliteit in Engeland in de 19e eeuw onder invloed van industrialisatie en verstedelijking. We trekken drie steekproeven van acht gemeenten: industriële, handels- en rurale gemeenten. We hebben gegevens over het aantal moorden per 10,000 inwoners. De vraag is nu: zijn de verschillen in het moordniveau tussen de verschillende typen van gemeenten statistisch significant? We stellen de hypothese dat verstedelijking criminaliteit in de hand werkt omdat de anonimiteit in steden groter is dan op het platteland.

Variantieanalyse geeft antwoord op de vraag of de nominale variabele 'type gemeente' (industrieel, commercieel, ruraal) statistisch significant van invloed is op de ratiovariabele 'moordniveau'.

Onder de aanname dat de standaardafwijkingen in de populatie gelijk zijn, kun je op basis van de steekproeven de variantie op twee manieren schatten:

- Het gewogen gemiddelde van de variantie binnen iedere groep gemeenten: de **binnengroepsvariatie (within-groups)**.
- De variantie van de gemiddelden van de drie groepen gemeenten rondom het algemeen gemiddelde: de **tussengroepsvariatie (between-groups)**.

Als de gemiddelden in de populatie hetzelfde zijn, leveren beide berekeningswijzen een identiek getal op. Als de gemiddelden niet hetzelfde zijn, zal de tweede schatting een groter getal opleveren dan de eerste. **Hoe groter de spreiding tussen de groepen ten opzichte van de spreiding binnen de groepen, hoe meer de groepen onderling verschillen en hoe sterker het verband tussen de nominale en de interval variabele.**

Bij variantieanalyse worden beide schatters van de variantie in de populatie vergeleken door ze op elkaar te delen. Als de binnengroeps- en de tussengroepsvariatie gelijk zijn, is de breuk, de **F-ratio**, gelijk aan 1. Als de tussengroepsvariantie groter is dan de binnengroepsvariatie, is de breuk groter dan 1. Naarmate de verschillen in de gemiddelden tussen de groepen groter zijn in vergelijking met de verschillen binnen iedere groep, is de breuk groter.

Natuurlijk kan het verschil toevallig zijn, maar hoe groter F (het verschil tussen de varianties) is, hoe kleiner de kans dat dit door het toeval komt. Bij iedere F-waarde kan de kans (p-waarde, *probability*) dat deze door het toeval is bepaald, berekend worden. Als de p-waarde kleiner is dan 0,05 is de kans dat de waarde aan het toeval te wijten is, kleiner dan 5%. Het verschil tussen de gemiddelden is dan significant, ofwel: de nominale variabele heeft een significante invloed op de interval-variabele.

### ***De berekening van de F-ratio***

De berekening van de variaties vindt plaats aan de hand van de som van de gekwadrateerde afwijkingen van het gemiddelde (Engels: *Sum of Squares*, SS). De opdeling van de totale variatie in een tussengroepsvariatie is niet nieuw. Een soortgelijke redenering hebben we eerder behandeld bij de regressieanalyse, met name toen we spraken over de determinatiecoëfficiënt en de opdeling van de totale variantie in een afhankelijke variabele in verklaarde variantie en onverklaarde variantie.

$$SS_{Total} = SS_{Groepen} + SS_{Error}$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{ij})^2$$

$$\text{Total SS} = \text{Within Groups SS} + \text{Between Groups SS}$$

*Within Groups SS* = som van de gekwadrateerde afwijkingen van de individuele scores van hun eigen groepsgemiddelde: de 'niet-verklaarde' afwijkingen.

*Between Groups SS* = som van de gekwadrateerde afwijkingen van de groepsgemiddelden van het algemene gemiddelde: de 'verklaarde' afwijkingen.

In het ideale geval (perfect verband) is alle variatie toe te schrijven aan de verschillen tussen de groepen en niets aan het verschil binnen de groepen. Je krijgt de geschatte variantie door de gekwadrateerde afwijkingen te delen door het aantal vrijheidsgraden (Engels: *Degrees of Freedom, DF*).

$$\text{Total } DF = \text{Within } DF + \text{Between } DF$$

$$\frac{\text{Verklaarde variantie}}{\text{Niet-verklaarde variantie}} = \frac{\text{Between } SS / DF}{\text{Within } SS / DF}$$

## 15. Zelf uitrekenen van een variantieanalyse

Moorden per 10.000 inwoners in 24 Engelse gemeenten (19<sup>de</sup> eeuw)

	Homicide rate Industriesteden	Homicide rate Handelssteden	Homicide rate Agrarische nederzettingen	Totaal
	4,3	5,1	12,5	
	2,8	6,2	3,1	
	12,3	1,8	1,6	
	16,3	9,5	6,2	
	5,9	4,1	3,8	
	7,7	3,6	7,1	
	9,1	11,2	11,4	
	10,2	3,3	1,9	
<b>Som</b>	<b>68,6</b>	<b>44,8</b>	<b>47,6</b>	<b>161</b>
<b>Gemiddelde</b>	<b>8,575</b>	<b>5,6</b>	<b>5,95</b>	<b>6,708333</b>
<b>N</b>	<b>8</b>	<b>8</b>	<b>8</b>	<b>24</b>

Hieronder illustreren we hoe een variantieanalyse dient uitgerekend te worden. We zien in het voorbeeld 24 steden. Er zijn 8 industriesteden, 8 handelssteden en 8 agrarische nederzettingen.  $3 \times 8 = 24$  steden. Nu kan je deze tabel bekijken en al met het blote oog vaststellen dat deze drie groepen verschillen van elkaar: de gemiddelden verschillen van elkaar. Als we geen rekening houden met deze typologie, dan is het grote gemiddelde gelijk

aan 6.7 moorden per 10.000 inwoners. Echter, voor de industriesteden is dit gemiddelde 8.575 moorden per 10.000 inwoners. Voor handelssteden bedraagt het gemiddelde 5.6 moorden per 10.000 inwoners en voor agrarische nederzettingen bedraagt het gemiddelde 5.95 moorden per 10.000 inwoners. Er zijn wel degelijk verschillen, zo veel is duidelijk. De vraag alleen is natuurlijk: hoe significant zijn deze verschillen. Kunnen we zeggen dat deze groepen statistisch significant van elkaar verschillen? Of zijn deze verschillen louter aan het toeval te wijten? **We zetten ons even aan het rekenen om dit te weten te komen. Welke informatie hebben we nodig? We moeten voor elke groep van steden inzicht krijgen in de variabiliteit rond het eigen groepsgemiddelde en we moeten inzicht krijgen in de verschillen tussen de drie groepen. Als de verschillen tussen de groepen groter is dan de verschillen tussen de steden, dan is er een significant verschil.**

**Dit heb je nodig:**

- **SS: Sum of Squares** = som van gekwadrateerde afwijkingen tegenover het gemiddelde
- **DF: Degrees of Freedom** = vrijheidsgraden = aantal waarnemingen of groepen - 1 (bij steekproeven)
- **Between Groups SS**: Tussengroepsvariatie of variatie tussen de gemeentetypes
- **Within Groups SS**: Binnengroepsvariatie of variatie binnen de gemeentetypes
- **Total SS** = **Between Groups SS** + **Within Groups SS**  
**Total DF** = **Between Groups DF** + **Within Groups DF**
- **Mean Squares** = Variantie = SS / DF

**Variantie tussen de groepen** = 'verklaarde variantie' in de afhankelijke variabele (Engels: *Dependent*) door de onafhankelijke variabele (Engels: *Factor*); wordt soms ook '*Model*' genoemd in SPSS.

**Variantie binnen de groepen** = 'niet-verklaarde variantie'; wordt ook '*Error*' genoemd in SPSS.

**F-Ratio** = **variantie tussen groepen / variantie binnen groepen** (= **verklaarde / niet-verklaarde variantie**). Dus: hoe hoger de F, hoe groter de verschillen tussen de groepen in verhouding tot de verschillen binnen de groepen.

**1)** We beginnen onze analyse met de studie van de variatie binnen elke groep uit te rekenen.  
We beginnen met de industriesteden

**a) Binnengroepsvariatie voor de *industriesteden*:**

$$(\bar{x} = 8.57)$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
4.3	-4.27	18.23
2.8	-5.77	33.29
12.3	3.73	13.91
16.3	7.73	59.75
5.9	-2.67	7.13
7.7	-0.87	0.76
9.1	0.53	0.28
10.2	1.63	2.66
<b>variatie=</b>		136.01

We starten met de individuele afwijkingen tegenover het gemiddelde en kwadrateren in de kolom daarnaast en maken de som. De variatie voor de industriesteden bedraagt 136.01.

**b) Binnengroepsvariatie voor de *handelssteden*:**

$$(\bar{x} = 5.6)$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
5.1	-0.5	0.25
6.2	0.6	0.36
1.8	-3.8	14.44
9.5	3.9	15.21
4.1	-1.5	2.25
3.6	-2	4
11.2	5.6	31.36
3.3	-2.3	5.29
<b>variatie=</b>		73.16

Nu berekenen we volledig op dezelfde manier de variaties voor de handelssteden. We starten met de individuele afwijkingen tegenover het gemiddelde en kwadrateren in de kolom daarnaast en maken de som. Deze bedraagt 73.16.

c) **Binnengroepsvariatie voor de *agrarische nederzettingen***

$$(\bar{x} = 5.95)$$

<b>x</b>	<b><math>x - \bar{x}</math></b>	<b><math>(x - \bar{x})^2</math></b>
12.5	6.55	42.90
3.1	-2.85	8.12
1.6	-4.35	18.92
6.2	0.25	0.06
3.8	-2.15	4.62
7.1	1.15	1.32
11.4	5.45	29.70
1.9	-4.05	16.40
<b>variatie=</b>		122.04

Tenslotte berekenen we volledig op dezelfde manier de variaties voor de agrarische nederzettingen. We starten met de individuele afwijkingen tegenover het gemiddelde en kwadrateren in de kolom daarnaast en maken de som. Deze bedraagt 122.04.

Nu we de drie afzonderlijke variaties hebben uitgerekend, kunnen we de totale binnengroepsvariatie berekenen. Dit is niet meer dan de optelsom van de afzonderlijke binnengroepsvariaties. De totale binengroepsvariantie is de totale binnengroepsvariatie gedeeld door het aantal vrijheidsgraden. Het aantal vrijheidsgraden is hier 21. Dat is het gevolg van het feit dat er in elke groep een element niet kon variëren aangezien elk groepsgemiddelde vastlag. Aangezien we drie groepen hebben, bedraagt het aantal vrijheidsgraden hier 24-3 en dus 21.

**De totale binnengroepsvariantie =**

$$\text{binnengroepsvariatie a} + \text{binnengroepsvariatie b} + \text{binnengroepsvariatie c) / aantal vrijheidsgraden} = \frac{(136.01+73.16+122.04)}{(24-3)} = 15.77$$

**2) Variatie tussen de groepen:**

Om de variatie tussen groepen te kennen moeten we ons baseren op de afwijkingen van elk groepsgemiddelde tegenover het grote gemiddelde. We zien een formule met een teller en een noemer. In de teller tellen we de gekwadrateerde verschillen op, dit wil zeggen we tellen de verschillen op tussen elk groepsgemiddelde en het grote gemiddelde, vermenigvuldigd met het aantal eenheden in elke groep. Omdat het aantal eenheden nu gemakshalve even groot is voor de drie groepen, kunnen we die groepsgrootte buiten de haakjes plaatsen.

In de noemer komt het aantal vrijheidsgraden. Er zijn drie groepen, en dat wil zeggen dat er twee groepen vrij kunnen variëren. Het aantal vrijheidsgraden is immers  $n-1$ , dus  $3-1 = 2$ . De idee is dat er altijd één eenheid geen vrij variërende waarde kan hebben als het gemiddelde vastligt.

$$8 * [( \text{gemiddelde industriesteden} - \text{totaal gemiddelde})^2 + (\text{gemiddelde handelssteden} - \text{totaal gemiddelde})^2 + (\text{gemiddelde agrarische nederzettingen} - \text{totaal gemiddelde})^2] / \text{aantal vrijheidsgraden}$$

**De tussengroepsvariantie bedraagt**

$$\frac{8 * [(8.57 - 6.71)^2 + (5.6 - 6.71)^2 + (5.95 - 6.71)^2]}{3-1} = \frac{8 * (3.46 + 1.23 + 0.58)}{2} = 21.08$$

We zien in elk geval in deze analyse dat de verschillen tussen de groepen groter zijn dan de verschillen binnen de groepen. Maar hoe significant zijn deze verschillen? De F-toets leert het ons.

**Laat ons eerst de F-waarde berekenen: F is de verhouding tussen de tussengroepsvariantie (21.08) en de totale binnengroepsvariantie (15.77). Deze breuk is heel eenvoudig te berekenen.  $F = 21.08 / 15.77 = 1.34$ .**

Opgelet! De gevonden F-waarde dien je nu op te zoeken in de tafel van F-waarden en je moet de waarde vergelijken met de kritische F-waarde. Wat is de kritische F-waarde? Dat is de waarde die past bij een gegeven aantal vrijheidsgraden in teller en noemer en een bepaald significantieniveau. Je F-waarde moet groter zijn dan de kritische gegeven een bepaald significantieniveau en vrijheidsgraden in de teller en noemer. Deze tabel vind je in bijlage. Op basis daarvan kan besloten worden of de gemiddelde scores tussen groepen significant van elkaar verschillen.

We hebben 21 vrijheidsgraden in de noemer en 2 vrijheidsgraden in de teller. Dit zoeken we op in de F tabel bij het 95% significantieniveau. Onze F waarde ( $F = 1.34$ ) is duidelijker lager dan de kritische waarde 3.47. Het verdict is duidelijk: het door ons onderzochte verband is NIET statistisch significant. We kunnen de nulhypothese dat de steden niet verschillen van het algemene gemiddelde niet verwerpen.

### **Een determinatiecoëfficiënt voor de variantieanalyse**

In een variantieanalyse bestaat er een equivalent voor de determinatiecoëfficiënt uit een regressieanalyse, met name **eta-kwadraat**. Eta-kwadraat is steeds *de verhouding tussen de tussengroepsvariatie en de totale variatie in Y*.

Toegepast op het voorbeeld bekomen we volgende waarde voor eta-kwadraat:

$42.16/373.33 = 0.1129 \rightarrow 11.29\%$  (tussengroepsvariatie/totale variatie. Totale variatie = tussengroepsvariatie PLUS binnengroepsvariatie of  $42.16 + 331.17$ ).

## **16. Voorbeelden van statistische inferentie in andere analysetechnieken**

We hebben verschillende analysetechnieken besproken en bij elke analyse hebben we aandacht gehad voor de beschrijvende resultaten. In dit hoofdstuk is het tijd om nu eens te kijken naar de interpretatie van de statistische significantie van de beschrijvende analyses die werden gepresenteerd.

## De toets op significantie van regressieparameters

		Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.613	.039		41.488	.000	1.537	1.689
	cumulative negative events	.455	.040	.184	11.319	.000	.376	.533

a. Dependent Variable: totale frequentie van regelovertredend gedrag

We bespreken eerst de regressiecoëfficiënten, i.e. de parameters van de best passende rechte, het intercept en de richtingscoëfficiënt en daarna de significantietoets van de determinatiecoëfficiënt. Hierboven presenteren we de resultaten van een regressieanalyse waarbij de afhankelijke variabele het totale aantal zelf-gerapporteerde delicten is en de onafhankelijke variabele het aantal negatieve levensgebeurtenissen. Sommige theorieën legden nogal veel klemtoon op de rol van negatieve levensgebeurtenissen (scheiding ouders, verlies of dood van een partner of familielid, ...). Deze theorieën benadrukt dat het aantal negatieve levensgebeurtenissen waaraan men blootgesteld wordt een goede predictor is voor de frequentie van criminaliteit. Welnu, in de internationale zelfrapportagestudie hebben we de kans gehad om deze veronderstelling na te gaan en de resultaten zie je in de tabel hierboven. De afhankelijke variabele is het aantal regelovertrledingen (totale frequentie de afgelopen 12 maanden) en de onafhankelijke variabele is het totaal aantal negatieve gebeurtenissen. De tabel geeft het intercept weer en de richtingscoëfficiënt. We hebben eerder reeds behandeld hoe je die coëfficiënten dient te interpreteren, maar nu gaan we in op de statistische significantietoets. Hiertoe is het belangrijk de juiste informatie te lezen. De tabel bevat naast het intercept en de richtingscoëfficiënt ( $B_0$  en  $B_1$ ) ook de standaardfout voor de betrokken parameter. Hoe weten we nu of de waarde die we in onze steekproef hebben gekregen ook statistisch significant is? We kunnen dit gemakshalve aflezen uit de laatste kolom, waar "sig" de afkorting is voor significantie en waar je het exacte significantieniveau kan aflezen. Je ziet dat beide coëfficiënten statistisch significant verschillen van nul. Op basis van de ongestandaardiseerde parameters en de standaardfout (de standard error) kan de t-waarde berekend worden. De t-waarde moet een waarde hebben die groter is dan 1.96 om te kunnen zeggen dat een verband statistisch significant is. Je moet die t-waarde niet met de hand zelf kunnen uitrekenen, maar je dient deze wel te kunnen interpreteren. Deze t-waarde kan opgezocht worden in de tabel in bijlage en de t-waarde dient eveneens groter te zijn dan de

kritische t-waarde die je kan opzoeken voor elk gewenst significantieniveau en aantal vrijheidsgraden. Hier is de t-waarde duidelijk significant. Maar dat wist je al omdat we er op wezen dat t-waarden die groter zijn dan 1.96 zeker significant zijn op het 0.05 niveau.

Vervolgens kijken we naar de analyse van de determinatiecoëfficiënt.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.184 <sup>a</sup>	.034	.034	2.349

a. Predictors: (Constant), cumulative negative events

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	706.673	3653	706.673	128.113	.000 <sup>b</sup>
	Residual	20150.038		5.516		
	Total	20856.711				

a. Dependent Variable: **totale frequentie van regelovertredend gedrag**

b. Predictors: (Constant), cumulative negative events

De determinatiecoëfficiënt bedraagt slechts 3.4%. Dit wil zeggen dat slechts 3.4 procent van de individuele verschillen in delinquent gedrag kan verklaard worden door de negatieve levensgebeurtenissen. Dat is heel weinig, maar toch verschillend van nul. IS dit nu significant verschillend van nul? Ja, zo blijkt, want de ANOVA tabel bevat de *regression sum of squares* en *residual sum of squares* met het gegeven aantal vrijheidsgraden. Voor de regression sum of squares is er maar één vrijheidsgraad omdat er slechts een onafhankelijke variabele in het model betrokken is. De residual sum of squares bedraagt 3653. Je kan de breuk berekenen: tussen elke mean square en dan bekom je de F waarde. Deze F-waarde kan je opnieuw gaan opzoeken in je kritische tabel, met het aantal vrijheidsgraden in teller en noemer, en dan merk je dat de determinatiecoëfficiënt significant verschilt van nul.

### De toets op significantie van een correlatiecoëfficiënt

Alle correlatiecoëfficiënten (Pearson, Spearman, Kendal, Gamma, ...) worden op hun significantie getoetst in SPSS. De output vermeldt steeds de exacte p-waarde. Als je een p-waarde tegenkomt die lager is dan 0.05 is het verband significant op het niveau 0.05. Doorgaans rapporteert men ook het niveau 0.01 en 0.001. Is een verband echter hoger is dan

0.05 is het verband strikt genomen niet meer statistisch significant. De vraag die zich stelt, en waarop het perfecte antwoord niet bestaat, is de vraag wat te doen bij een randgeval. Ook op nominaal niveau vind je in de output van SPSS de exacte significantie voor parameters zoals Phi en Cramer's V en Chi-kwadraat. De interpretatie is telkens analoog.

### Randsignificantie en bedrog?

Wetenschappers worden vaak geconfronteerd met niet-perfecte data en resultaten. Als een p-waarde nu 0.06 bedraagt. Wat doe je dan? De conventie wil dat we verbanden tussen 0.05 en 0.10 als **randsignificant** gaan beschouwen. We toetsen immers steeds tweezijdig en vaak is men vrij zeker van een verband. Als je bijvoorbeeld een verband wil onderzoeken tussen de sociale bindingen van iemand en de criminaliteit die iemand pleegt, dan is de hypothese normaal gezien dat hoge mate van sociale bindingen gepaard gaat met lagere betrokkenheid bij criminaliteit. Echter, dat is niet steeds zo. Dus: voorzichtigheid is geboden. Je zal wellicht al in de krant gelezen hebben dat sommige onderzoekers betrapt werden op het mooier maken van de gegevens. Dat is puur bedrog en gebeurt om significante resultaten nog significanter te maken. Bij mijn weten is het in de criminologie nog niet gebeurd, maar in de farmacie is het al gebeurd dat onderzoekers hun p-waarde hebben opgelapt, speciaal om de tests er beter te laten uitzien. Waarom zou men dat doen? Nou, als er veel geld mee gemoeid is, bijvoorbeeld een nieuw medicijn tegen een moeilijk te overwinnen ziekte, dan is het voor diegene die het medicijn heeft uitgevonden wel relevant om het op de markt te krijgen. Met een randsignificante p-waarde wordt zoiets alvast moeilijk. Zo zie je maar: het is niet de statistiek die liegt, maar de gebruiker.

## 17. Testvragen

Hieronder vinden Je enkele uitspraken over de bivariate statistiek en inferentiële statistiek. Deze vragen kan je gebruiken om je parate kennis te toetsen over de basiskennis die tot nog toe werd meegegeven. **In gewijzigde vorm kunnen dergelijke theorievragen ook op het examen voorkomen.** Deze vragen zijn afkomstig uit vroegere examens. De correcte antwoorden zijn rechtstreeks uit de cursus afleidbaar en worden vanuit didactisch oogpunt niet meegegeven. Dit is niet meer dan een test om te zien of je mee bent met de leerstof. Als je op deze test faalt, dan is het hoogdringend tijd om in actie te schieten.

**1. De Y-as wordt ook wel het ordinaat genoemd**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**2. De X-as wordt ook wel de abscis genoemd**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**3. Een scatterplot kan gebruikt worden om de relatie tussen een ratio-variabele en een interval-variabele te presenteren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**4. Een scatterplot kan gebruikt worden om de relatie tussen een ordinale variabele en een interval-variabele te presenteren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**5. Een spreidingsdiagram kan gebruikt worden om de relatie tussen een nominale en een ordinale variabele te presenteren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**6. Het  $(x,y)$  coördinaat van het rekenkundig gemiddelde van  $x$  en het rekenkundig gemiddelde van  $y$  is het bivariate zwaartepunt van het spreidingsdiagram**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**7. De ongestandaardiseerde regressiecoëfficiënt (b1) uit een bivariate regressieanalyse is een**

- Symmetrische maat
- Asymmetrische maat

**8. De bivariate gestandaardiseerde regressiecoëfficiënt uit de regressie van Y op X is gelijk aan de gestandaardiseerde covariantie tussen x en y**

- Deze uitspraak is juist
- Deze uitspraak is fout

**9. De regressie van Q op S wil zeggen dat we**

- Q als afhankelijke variabele hebben en S als onafhankelijke variabele
- S als afhankelijke variabele hebben en Q als onafhankelijke variabele

**10. Het percentageverschil is een symmetrische associatiemaat**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**11. Hierna vind je een tabel waarbij de relatie tussen twee nominale variabelen, X en Y wordt voorgesteld. We veronderstellen dat R de afhankelijke variabele is en dat S een causale invloed uitoefent op R.**

	Variabele S			Totaal
Variabele R	A	B	E	
	C	D	F	
Totaal	G	H	I	

Welke van de uitspraken is juist:

- E en F noemen we kolommarginalen
- E en F noemen we rijmarginalen (ik twijfel!)
- A / G geeft de proportie van A
- A / C geeft de odds op A

- Het juiste percentageverschil moet berekend worden op basis van de vergelijking van de verschillen tussen

**12. Chi-kwadraat is een associatiemaat die zeer gevoelig is aan N**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**13. Als je N vermenigvuldigt met 2, dan wordt de waarde van chi-kwadraat twee keer zo groot**

- Deze uitspraak is juist
- Deze uitspraak is fout

**14. Chi-kwadraat moet worden berekend op basis van de ruwe scores**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**15. Chi-kwadraat kan ook worden berekend op basis van de proporties**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**16. In een 2\*2 tabel is de waarde van Phi gelijk aan de waarde van V**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

### 17. Bekijk volgende tabel

Slachtofferschap vijf jaar	afgelopen	Scholingsgraad
Nooit	A      B      C	Laag      Midden      Hoog
Een maal	D      E      F	
Twee maal of meer	G      H      I	

### Welke uitspraken over deze tabel zijn juist

- $(E^*F)$  is een consistent paar
- $(F^*H)$  is een inconsistent paar
- Gamma is gebaseerd op de verhouding tussen de consistentie paren min de inconsistentie paren en de consistentie paren plus de inconsistentie paren
- Gamma is een associatiemaat die monotoniciteit veronderstelt (weet ik niet of dit nu ook nog juist is of niet)
- Gamma is een asymmetrische maat (en symmetrisch)
- Gamma wordt gebruikt op nominaal niveau (zeker ordinaal)

### 18. Welke uitspraken zijn juist?

- Spearman's Rho is een rangcorrelatiecoëfficiënt
- Pearson's r is een rangcorrelatiecoëfficiënt (= product-moment correlatiecoëfficient)
- Kendall's Tau- is een rangcorrelatiecoëfficiënt
- Rangcorrelaties zijn asymmetrische maten (symmetrisch)

### 19. Welke uitspraken over Pearson's r zijn juist?

- De bivariate correlatiecoëfficiënt van Pearson veronderstelt lineariteit (zoniet kan Pearson's r leiden tot verkeerde conclusies)
- De bivariate correlatiecoëfficiënt is een gestandaardiseerde covariantie
- De bivariate correlatiecoëfficiënt wordt berekend uit niet gekwadrateerde deviatiescores

**20. De bivariate correlatiecoëfficiënt tussen ouderlijk toezicht en criminaliteit en y is negatief en de gestandaardiseerde rico uit de regressie van criminaliteit op toezicht bedraagt -0.35.**

- Daaruit volgt dat toezicht een remmend effect heeft op criminaliteit
- Daaruit volgt dat toezicht nefast is voor de criminaliteit want meer toezicht, resulteert in meer criminaliteit
- Daaruit volgt dat meer toezicht ongerelateerd is aan criminaliteit

**21. De covariatie is**

- De kruisproductensom
- De som van de deviatiescores van x + de deviatiescores van y

**22. Hieronder volgen een aantal uitspraken over de lineaire regressieanalyse**

- De lineaire regressie is een asymmetrische techniek
- De lineaire regressie (het basismodel) is gebaseerd op de kleinste kwadratenoplossing
- De lineaire regressie is niet geschikt voor kwadratische curvilineaire relaties (de andere twee weet ik niet zo goed, maar in fb document staan ze alle 3 wel aangeduid als juist)

**23. Het residu is**

- Het verschil tussen geobserveerde en verwachte waarde
- Het verschil tussen verwacht en geobserveerde waarde

**24. De som van de gekwadrateerde residuelen bedraagt nul**

- Deze uitspraak is juist
- Deze uitspraak is fout

**25. De variatie in Y kan ook uitgedrukt worden als de som van de regressie sum of square minus de residual sum of square**

- Deze uitspraak is juist
- Deze uitspraak is fout (niet minus, maar plus!)

**26. Als de punten uit een puntenwolk heel dicht bij de regressierechte liggen, dan kunnen we vermoeden dat**

- De model fit zeer hoog zal zijn
- De model fit zeer laag zal zijn

**27. Het is perfect mogelijk dat twee regressiecoëfficiënten een zelfde beta-waarde hebben, maar een verschillende aliënatiocoëfficiënt hebben**

- Deze uitspraak is juist
- Deze uitspraak is fout

**28. De variantieanalyse wordt gebruikt wanneer**

- De onafhankelijke variabele nominaal is en de afhankelijke variabele metrisch
- De afhankelijke variabele nominaal is en deon afhankelijke variabele metrisch

**29. Eta-kwadraat komt qua interpretatie overeen met een determinatiecoëfficiënt**

- Deze uitspraak is juist
- Deze uitspraak is fout

**30. De toets van de significantie van de determinatiecoëfficiënt gebeurt aan de hand van**

- De F-toets
- De T-toets

**31. De toets van het intercept en de rico gebeurt aan de hand van de**

- De F-toets
- De T-toets
- De Z-toets

**32. Intervalschatting is een centraal element uit de**

- Inferentiële statistiek
- Descriptieve statistiek

**33. De steekproevenverdeling van X is**

- De verdeling van X
- De verdeling van alle gemiddelde waarden voor X in een reeks van steekproeven met gelijke omvang
- De verdeling van alle gemiddelde waarden voor X in een reeks van steekproeven met ongelijke omvang
- Geen van de voorgaande beweringen is juist

**34. Een betrouwbaarheidsinterval van met een alfa (kans op vergissing) van 10 %**

= Z-score van 90

- Komt overeen met een z-score van 1.645
- Komt overeen met een z-score van 1.960
- Komt overeen met een z-score van 2.576

**35. Ik doe criminologisch onderzoek naar belastingontduiking en stel vast dat het gemiddeld aantal veroordelingen onder belastingontduikers (mean = 3) lager is dan onder inbrekers (mean = 5). De betrouwbaarheidsintervallen van de beide gemiddeldes blijken elkaar te overlappen.**

Hieruit besluit ik dat

- De gemiddeldes niet significant van elkaar verschillen
- De gemiddeldes wel significant van elkaar verschillen

**36. Een type I fout betekent :**

- De nulhypothese is juist en ik verworp ze foutief
- De nulhypothese is verkeerd en ik behoud ze foutief

**37. Een type II fout betekent :**

- De nulhypothese is juist en ik verworp ze foutief
- De nulhypothese is verkeerd en ik behoud ze foutief

**38. De power van een test berekenen is belangrijk want hierdoor houdt men rekening met zowel type I als type II fouten**

- Deze uitspraak is juist
- Deze uitspraak is fout (de power van de test is de kans dat een toets geen type-II fout maakt of de kans op het terecht verwerpen van een foute nulhypothese)

**39. De tussengroepsvariantie is de tussengroepsvariatie gedeeld door het aan tal vrijheidsgraden tussen groepen**

- Deze uitspraak is juist
- Deze uitspraak is fout

**40. Als de tussengroepsvariantie groter is dan de binnengroepsvariantie, en de F-toets geeft een waarde die hoger is dan de kritische waarde, is het verband tussen x en y significant**

- Deze uitspraak is juist
- Deze uitspraak is fout

**41. Als de tussengroepsvariantie groter is dan de binnengroepsvariantie, en de F-toets geeft een waarde die lager is dan de kritische waarde, is het verband tussen x en y significant**

- Deze uitspraak is juist
- Deze uitspraak is fout

## **18. Leerdoelen**

Dit deel van het handboek behandelt de praktijk van het schatten en toetsen, twee uiterst belangrijke procedures in de inferentiële statistiek. Criminologen doen hierop heel vaak beroep wanneer zij steekproeven trekken en bevolkingsbevragingen organiseren omtrent slachtofferschap, onveiligheidsbeleving, attitudes tegenover het strafrechtssysteem,... Het is van belang de uitkomsten van steekproeven naar waarde te schatten en dit kan via de principes van schatten en toetsen. De centrale begrippen die verband houden met de inferentiële statistiek, zoals schatten, toetsen, p-waarden, nulhypothese, alternatieve hypothese, het betrouwbaarheidsinterval, variantieanalyse, F-waarden dienen te kunnen worden toegepast. De steekproevenverdeling werd geïntroduceerd in dit hoofdstuk als een bijzondere verdeling van steekproefuitkomsten (gemiddelden, standaardafwijkingen, correlaties,...).

Het principe van de centrale limietstelling dient zeer goed begrepen te worden. In de oefensessies voorzien we hierin simulaties om deze abstracte kennis in de praktijk toe te passen. We verwachten dat studenten betrouwbaarheidsintervallen ook zelf met de hand kunnen uitrekenen en dat deze geïnterpreteerd kunnen worden wanneer resultaten van wetenschappelijk onderzoek onder de vorm van output van statistische verwerkingspakketten worden voorgelegd aan studenten. Studenten worden tijdens de oefensessies getraind om vragen te beantwoorden aan de hand van output van het verwerkingspakket SPSS. Het is verder van belang te weten dat bij het toetsen niet enkel de normale verdeling wordt gebruikt, maar ook andere verdelingen. De principes zijn echter dezelfde. Op basis van deze verdelingen kunnen we kansen nagaan dat een bepaalde parameter in de populatie waaruit de steekproef afkomstig is effectief bestaat. Ook de variantieanalyse dient door de studenten zelf uitgerekend te kunnen worden. Hieronder voorzien we nog in een samenvattende tabel van toetsingsprocedures voor associaties categorisch en metrisch niveau.

**Samenvattende tabel: inferentie op de geziene associatiematen en analysetechnieken**

<b>VERBANDEN TUSSEN 2 VARIABELEN: niet-dependente associatiematen (samenvang)</b>			
	<b>Nominaal</b>	<b>Ordinaal</b>	<b>Metrisch</b>
<b>verbanden tussen 2 variabelen</b>	Controleer de Chi-kwadraat toets op Phi, Cramer's V	Controleer de significantieniveaus van Spearman's rho en gamma	Controleer het significantieniveau van de correlatiecoëfficiënt van Pearson

**Samenvattende tabel: inferentie op de geziene dependente analysetechnieken (causatie en predicitie)**

<b>Onafhankelijke variabele</b>	<b>Afhankelijke variabele</b>	<b>ANALYSETECHNIEK</b>	<b>Toetsen</b>
Interval/Ratio	Interval/Ratio	lineaire regressieanalyse	T-toets op de regressiecoëfficiënten F-toets op determinatiecoëfficiënt $P < 0.05$ of beter!
Nominaal	Interval/Ratio	Een-factor variantieanalyse	F-toets op Sum of Squares between Groups Sum of Squares within Group



## Hoofdstuk 10

### De partiële correlatie als introductie tot de multivariate statistiek

#### 1. Inleiding

Tot nu begaven we ons op het domein van de bivariate statistiek. Bivariate analyses zijn zinvol. Ze leren ons of twee kenmerken samen geobserveerd worden bij statistische onderzoekseenheden dan louter op basis van toeval kan verwacht worden. Bivariate analyses hebben echter belangrijke beperkingen. In de sociale verklarende wetenschappen zijn we vaak geïnteresseerd in de effecten van meerdere onafhankelijke variabelen op een afhankelijke variabele. Dat is in de criminologie niet anders. Het grote probleem in de bivariate analyse zit hem in het feit dat we uit een bivariate analyse weinig kunnen besluiten. Een bivariaat verband kan potentieel veroorzaakt worden door een derde, storende variabele waar we geen rekening mee hielden. Vaak omdat andere onafhankelijke variabelen, die aan eenzelfde afhankelijke variabele gerelateerd zijn, onderling ook samenhangen, is het bivariate effect vaak een overschatting van de realiteit. Dit is een zeer belangrijke reden om meer dan één onafhankelijke variabele in een analyse in te voeren. Van zodra we meerdere onafhankelijke variabelen in een analyse inbrengen, begeven we ons op het domein van de **multivariate statistiek**. Het meest éenvoudige model is het meervoudige *regressiemodel* met twee onafhankelijke variabelen. Dit is een eenvoudige extensie van de bivariate regressieanalyse. In de bivariate lineaire regressie onderzoeken we het effect van één onafhankelijke variabele op één afhankelijke variabele (onder de veronderstelling dat het verband tussen beide variabelen lineair is). Bij de meervoudige lineaire regressie onderzoeken we de effecten van meer dan één onafhankelijke variabele, maar nog steeds op slechts één afhankelijke variabele.<sup>16</sup>

Uiteraard is de lineaire wereld van de multivariate analyse niet beperkt tot het uitvoeren van een meervoudige regressieanalyse. De meervoudige lineaire regressieanalyse leent zich echter perfect om uiteen te zetten wat de beperkingen zijn van bivariate analyses. Waarom is de meervoudige analyse van statistische gegevens zo belangrijk in de sociale wetenschappen in het algemeen en de criminologie in het bijzonder? Hiervoor zijn verschillende redenen te bedenken.

---

<sup>16</sup> Zijn er meerdere afhankelijke variabelen, dan spreken we van *multivariate lineaire regressie*.

- **In de eerste plaats** is het zo dat de sociale werkelijkheid multivariaat is. Er zijn nu eenmaal **meerdere determinanten** verbonden aan regelovertredend gedrag. De criminoloog die ervoor kiest deze werkelijkheid aan de hand van statistische analysetechnieken te onderzoeken, dient zich hiervoor te wenden tot de multivariate analyse.
- **Ten tweede** is het zo dat het in de criminologie moeilijk is om fenomenen **geïsoleerd** te bestuderen. In gecontroleerde wetenschappen is het wel mogelijk om fenomenen aan de hand van experimenten in geïsoleerde positie te bestuderen. Dit gaat slechts in beperkte mate in de sociale wetenschappen. Zowel de causale als niet causale analyse van criminaliteitsfenomenen maakt daarom gebruik van het niet experimentele onderzoeksdesign als alternatief. Dit is het principe van **de statistische controle**. Elke variabele die gecorrelateerd is met de onafhankelijke variabele en die meebepalend kan zijn voor de score op de afhankelijke variabele, is een storende variabele. We geven een voorbeeld uit onderzoek naar individuele verschillen in delinquent gedrag: we zijn geïnteresseerd in de impact van morele normen op individuele betrokkenheid bij criminaliteit. Delinquente normen kunnen samenhang met de opvoeding, de sociale controle in het gezin, de bindingen die jongeren hebben met de conventionele samenleving, de buurt waarin men opgroeit, de mate waarin jongeren in contact komen met criminele rolpatronen,... We kunnen ons bijvoorbeeld afvragen of het verband tussen delinquentie waarden en individuele verschillen in delinquentie zou veranderen als we ook rekening zouden houden met ouderlijke controle.  
Controleren voor een storende variabele, bijvoorbeeld ouderlijke controle, betekent dat we de samenhang tussen delinquentie waarden en delinquent gedrag bestuderen voor jongeren uit gezinnen met eenzelfde niveau van ouderlijke controle. We houden de variabele ‘ouderlijke controle’ constant, zodat we alleen kijken naar variatie in de afhankelijke variabele ‘delinquent gedrag’ en in de onafhankelijke variabele ‘delinquentie waarden’ die niet samenhangt met een verschil in ouderlijke controle.
- **Ten derde** is het zo dat achter een bivariaat verband meer kan schuil gaan dan op het eerste zicht lijkt. We leggen dit verderop gedetailleerd uit met een criminologisch voorbeeld aan de hand van het “*schijneffect*” of de *spurieuze verbanden, maar ook de conditionele causaliteit of de statistische interactie*. Een spurieus verband verwijst naar een correlatie tussen X en Y omdat beide variabelen afhangen van een derde variabele Z. We behandelen het ‘schijneffect’ grondig in het vervolg van dit hoofdstuk. Interactie betekent dat de samenhang tussen twee variabelen afhankelijk is van de waarden van

een andere variabele. Er is sprake van interactie wanneer het effect van een variabele X op een variabele Y verschilt naargelang de waarden van een variabele Z. Bijvoorbeeld: omgevingskenmerken kunnen een effect hebben op criminaliteit maar de sterke van het effect kan afhankelijk zijn van bepaalde individuele kenmerken. We behandelen statistische interactie grondig in hoofdstuk 11.

In de volgende paragrafen leggen we de theoretische achtergrond van de partiële correlatie uit. We tonen hoe je zelf handmatig een partiële correlatiecoëfficiënt berekent.

## 2. De partiële correlatiecoëfficiënt

De partiële correlatie is de correlatie tussen twee variabelen (X1 en Y), onder statistische controle van één of meerdere storende variabelen. De partiële correlatieanalyse wordt uitgevoerd wanneer we willen nagaan of :

- de samenhang tussen X1 en Y **spurieus** is.
- de samenhang tussen X1 en Y **indirect** is.

Hierna geven we twee criminologische voorbeelden.

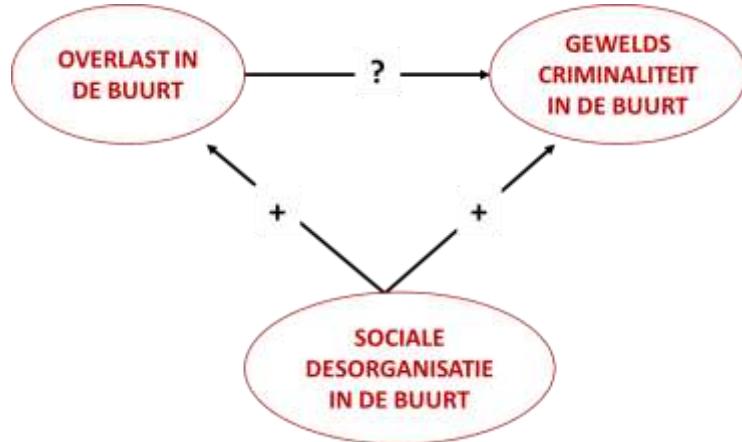
### *Spurieuze samenhang of een schijnverband: een voorbeeld uit de sociale desorganisatietheorie*

Het uitgangspunt betreft de samenhang tussen ‘overlast in de buurt’ en ‘geweldscriminaliteit in de buurt’. De vraag is of deze samenhang werkelijk kan beschouwd worden als een effectrelatie: als er meer overlast in de buurt is, leidt dit dan naar meer geweldscriminaliteit in de buurt ? Een alternatieve verklaring is immers eveneens mogelijk: namelijk een derde variabele kan verantwoordelijk zijn voor zowel ‘overlast in de buurt’ als voor ‘geweldscriminaliteit in de buurt’. In dat geval zou de samenhang tussen deze twee variabelen ‘wegverklaard’ worden door die derde variabele. De oorspronkelijk geobserveerde samenhang zou dan een schijnverband zijn.

Een voorbeeld van zo’n derde variabele is ‘sociale desorganisatie in de buurt’. Veronderstel een positieve relatie tussen sociale desorganisatie in de buurt en overlast in de buurt (hoe hoger de niveaus van sociale desorganisatie, hoe meer overlast) en veronderstel een positieve relatie tussen sociale desorganisatie in de buurt en geweldscriminaliteit in de buurt (hoe hoger de niveaus van sociale desorganisatie, hoe hoger de niveaus geweldscriminaliteit).

Figuur 1 geeft een visuele voorstelling van een spurieuze relatie of een schijnverband.

Figuur 1: Een spurieuze relatie: voorbeeld van een schijnverband uit de sociale desorganisatietheorie



De plustekens verwijzen naar een positieve samenhang, het vraagteken verwijst naar een potentieel spurieuze relatie. De bestaande bivariate samenhang is het gevolg van *een gemeenschappelijke oorzaak*. In het voorbeeld wordt sociale desorganisatie gezien als de gemeenschappelijke oorzakelijke factor die verklaart waarom buurten zowel te maken hebben met overlast als met geweldscriminaliteit.

In deze analyse zijn drie variabelen betrokken<sup>17</sup>, gemeten op intervalmeetniveau. De gemeenschappelijke oorzaak ‘sociale desorganisatie’ is de **controlevariabele**. De toegepaste analysetechniek is **de partiële correlatieanalyse**. De achterliggende veronderstelling is dat de oorspronkelijke samenhang tussen overlast en criminaliteit verdwijnt wanneer men **controleert** voor sociale desorganisatie. Concreet betekent dit dat buurten die verschillen in mate van overlast ook verschillen in mate van geweldscriminaliteit maar dat deze co-variatie volledig te wijten is aan verschillen in sociale desorganisatie. Voor buurten met gelijke niveaus van sociale desorganisatie, verdwijnt de samenhang tussen overlast en geweldscriminaliteit. De samenhang tussen overlast en geweldscriminaliteit is **spurieus of een schijnverband**.

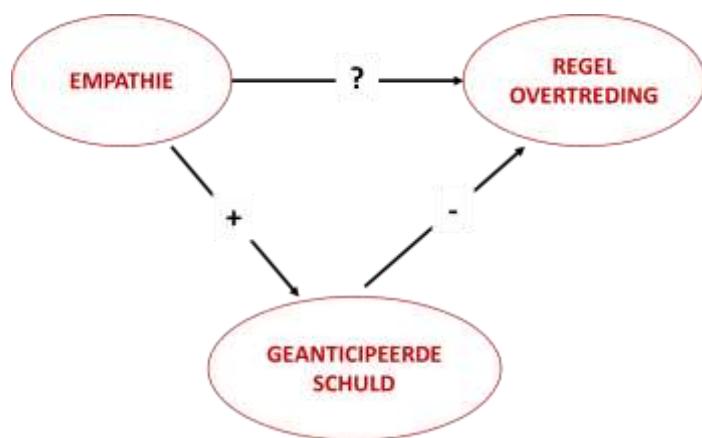
<sup>17</sup> Een partiële correlatieanalyse hoeft zich niet te beperken tot 3 variabelen ( $X_1$  en  $Y$  onder controle van  $X_2$ ). De analyse kan uitgebreid worden naar meerdere controlevariabelen. De partiële correlatie kan berekend worden tussen  $X_1$  en  $Y$  onder controle van  $X_2, X_3, X_4$ . In dit geval is de eerste stap in de analyse geen bivariate maar een multiple regressie analyse van  $X_1$  op  $X_2, X_3$  en  $X_4$ , hetgeen resulteert in de residuale term  $X_1 - X_1\hat{}$ . De tweede stap in de analyse is een multiple regressie analyse van  $Y$  op  $X_2, X_3$  en  $X_4$ , hetgeen resulteert in de residuale term  $Y - Y\hat{}$ . De zero-order correlatie tussen beide residuale termen is dan de partiële correlatiecoëfficiënt onder controle van  $X_2, X_3$  en  $X_4$ .

### **Indirecte relatie : toepassing van de empathie-ontwikkelingstheorie van Hoffman (2001)<sup>18</sup>**

Het uitgangspunt betreft de samenhang tussen ‘empathie’ en ‘regelovertreding’. De vraag is of deze samenhang werkelijk kan beschouwd worden als een effectrelatie: leidt meer/minder empathie tot minder/meer regelovertreding ? Het is immers mogelijk dat de samenhang te wijten is aan een vorm van indirecte samenhang via een intermediaire variabele. Een voorbeeld van zo’n intermediaire variabele is ‘geanticipeerde schuld’.

Veronderstel: we observeren een positieve relatie tussen empathie en geanticipeerde schuld (hoe meer empathie, hoe meer geanticipeerde schuld) en een negatieve relatie tussen geanticipeerde schuld en regelovertreding (hoe meer geanticipeerde schuld, hoe minder regelovertreding). Dit mechanisme van **indirecte samenhang** is visueel voorgesteld in Figuur 2.

**Figuur 2: Voorbeeld van een indirecte relatie gebaseerd op het empathie-ontwikkelingsmodel van Hoffman (2001)**



Het plus- en minteken verwijzen respectievelijk naar een positieve en negatieve samenhang. Het vraagteken refereert ernaar dat de oorspronkelijk veronderstelde samenhang tussen empathie en regelovertreding niet direct maar indirect van aard is. De toegepaste analysetechniek is ook in dit geval een **partiële correlatieanalyse**. Net zoals in het geval van spurieuze samenhang (zie hierboven) is de variantie die empathie (X1) en regelovertreding (Y) delen met elkaar, in het geval van indirecte samenhang, (bijna) volledig te wijten aan de variantie die beide variabelen delen met de controlevariabele ‘geanticipeerde schuld’. Dit betekent concreet dat voor respondenten met gelijke niveaus van geanticipeerde schuld, de

<sup>18</sup> Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

relatie tussen empathie en regelovertraving verdwijnt. Dit is ook een toepassing van het mechanisme van de partiële correlatie analyse: de samenhang tussen empathie en regelovertraving is **indirect**.

**Dus, het mechanisme van de partiële correlatieanalyse is hetzelfde voor zowel een spurieuze als indirecte samenhang. De richting van de pijlen in het conceptueel diagram is echter verschillend. In geval van:**

- **spurieuze relatie** is de controle variabele een gemeenschappelijke oorzaak (zie Figuur 1).  $(X_1 \leftarrow Z \rightarrow Y)$
- **Indirecte relatie** is de controle variabele een intermediaire variabele (zie Figuur 2).

$$(X_1 \rightarrow Z \rightarrow Y)$$

**Waarom is de berekening van een partiële correlatie nodig ?**

Waarom de berekening van een partiële correlatie nodig is, kunnen we best uitleggen aan de hand van een venndiagram.

Veronderstel dat we gegevens hebben over drie variabelen:

X = **Inspanningen** gedaan voor het studeren van het vak statistiek

Y = **De behaalde score** van de student op het examen statistiek

Z = De mate waarin de student **angst heeft voor het vak statistiek**

We vinden de volgende samenhangen :

X versus Y:  $r_{XY} = +0.20 \quad r^2_{XY} = 0.04$

X versus Z:  $r_{XZ} = +0.80 \quad r^2_{XZ} = 0.64$

Y versus Z:  $r_{YZ} = -0.40 \quad r^2_{YZ} = 0.16$

We zien een zwakke-matige positieve samenhang tussen inspanningen (X) en behaalde scores (Y) : hoe meer inspanningen, hoe hoger de scores ( $r_{YZ} = 0.20$ ).

We zien een zeer sterke positieve samenhang tussen inspanningen (X) en angst voor het vak (Z) : hoe meer inspanningen, hoe meer angst voor het vak ( $r_{XZ} = 0.80$ ).

We zien een matig-sterke negatieve samenhang tussen behaalde scores (Y) en angst voor het vak (Z) : hoe hoger behaalde scores, hoe lager angst voor het vak ( $r_{YZ} = -0.40$ ).

We constateren dus een samenhang tussen inspanningen en behaalde scores (tussen X en Y) maar ook tussen inspanningen en angst (tussen X en Z) en tussen behaalde scores en angst (tussen Y en Z).

We zien dat inspanningen (X) 4% van de variabiliteit deelt met behaalde scores (Y), 64% van de variabiliteit deelt met angst voor het vak (Z) en dat behaalde scores (Y) 16% van de variabiliteit deelt met angst voor het vak (Z).

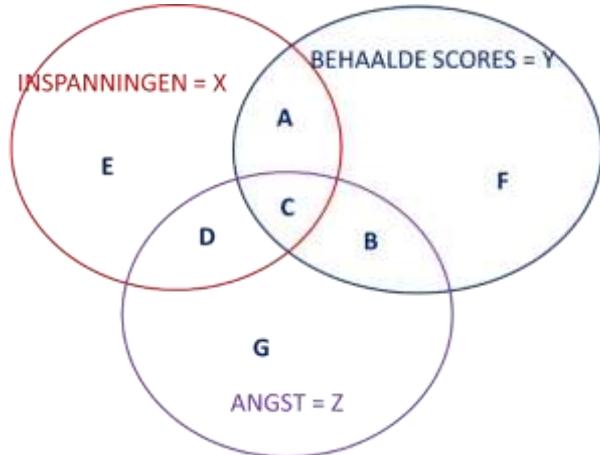
**Dus: inspanningen en behaalde scores (X en Y) delen variabiliteit of gemeenschappelijke variantie met angst voor het vak (Z).**

**Om een maat te krijgen van de unieke samenhang tussen inspanningen en behaalde scores moeten we rekening houden met angst voor het vak.**

Of anders geformuleerd: om een zuivere schatting te krijgen van de relatie tussen X en Y, moet de gemeenschappelijke variantie met Z, een “potentiële confounder” verwijderd worden. Die gemeenschappelijke variantie kan er immers toe leiden dat een bivariate analyse misleidende resultaten geeft.

Figuur 3 geeft een visuele voorstelling van de variantie van elke variabele en de gedeelde variantie.

*Figuur 3: Drie variabelen en de gedeelde variabiliteit*



We stellen X, Y en Z voor als cirkels. De volledige oppervlakte van elke cirkel is gelijk aan 100% van de variatie in elke variabele.

Figuur 3 toont tevens de overlap of variabiliteit die de variabelen met elkaar delen<sup>19</sup>:

<sup>19</sup> Om helemaal correct te zijn, zou de variatie die de variabelen met elkaar delen *proportioneel* moeten voorgesteld worden. Voorbeeld: als twee variabelen 10% van de variabiliteit gemeenschappelijk hebben, dan zou 10% van hun cirkeloppervlakte moeten ingekleurd worden als gemeenschappelijk. We hebben er in dit voorbeeld voor geopteerd om dit niet te doen teneinde de interpreteerbaarheid te vergroten.

- A = variabiliteit in behaalde scores (Y) **uniek** gedeeld met inspanningen (X)
- B = variabiliteit in behaalde scores (Y) **uniek** gedeeld met angst voor het vak (Z)
- C= variabiliteit in behaalde scores (Y) **gedeeld met inspanningen (X) en angst voor het vak (Z)**
- D= variabiliteit in inspanningen (X) **uniek** gedeeld met angst voor het vak (Z)
- E= variabiliteit in inspanningen (X) **niet** gedeeld met behaalde scores (Y) noch met angst voor het vak (Z)
- F = variabiliteit in behaalde scores (Y) **niet** gedeeld met inspanningen (X) noch met angst voor het vak (Z)
- G = variabiliteit in angst voor het vak (Z) **niet** gedeeld met inspanningen (X) noch met behaalde scores (Y)
- A + C = gedeelde variabiliteit tussen inspanningen en behaalde scores (X en Y) = 4%  
Waarvan A = unieke variabiliteit tussen inspanningen en behaalde scores  
Waarvan C = variabiliteit die inspanningen en behaalde scores delen met angst voor het vak
- C + B = gedeelde variabiliteit tussen behaalde scores en angst voor het vak (Y en Z) = 16%  
Waarvan B = unieke variabiliteit tussen behaalde scores en angst voor het vak  
Waarvan C = variabiliteit die behaalde scores en angst voor het vak delen met inspanningen
- C + D = gedeelde variabiliteit tussen inspanningen en angst voor het vak (X en Z) = 64%  
Waarvan D = unieke variabiliteit tussen inspanningen en angst voor het vak  
Waarvan C = variabiliteit die inspanningen en angst voor het vak delen met behaalde scores

Bedenk nu het volgende :

Gebied A = de variabiliteit in behaalde scores (Y) die **uniek** gedeeld wordt met inspanningen (X) maar in combinatie met C is dit de totale variatie in behaalde scores (Y) die gedeeld wordt met inspanningen (X) = 4%. De correlatie tussen behaalde scores (Y) en inspanningen (X) leert ons dat beide variabelen 4% van de variabiliteit delen. In werkelijkheid is maar een gedeelte uniek (gebied A) want er is een gedeelde overlap met angst voor het vak (gebied C).

Daarnaast hebben zowel inspanningen (X) als behaalde scores (Y) een unieke overlap met angst voor het vak (Z): respectievelijk gebied D en gebied B.

**De partiële correlatie is de unieke samenhang tussen twee variabelen X en Y onder controle van Z. Die unieke samenhang is de variatie in X en Y (gebied A) die overblijft als de volledige samenhang met Z is ‘weggenomen’ (dus gebieden B, C en D).**

De partiële correlatie analyse is een procedure die de gemeenschappelijke variantie verwijdert en de correlatie weergeeft tussen X en Y, onder de statistische controle van Z.

Het komt er dus op aan om de gemeenschappelijke variantie die X en Z, en Y en Z hebben op voorhand uit te schakelen, zodat wat overblijft de waargenomen variantie is in X en de waargenomen variantie in Y.

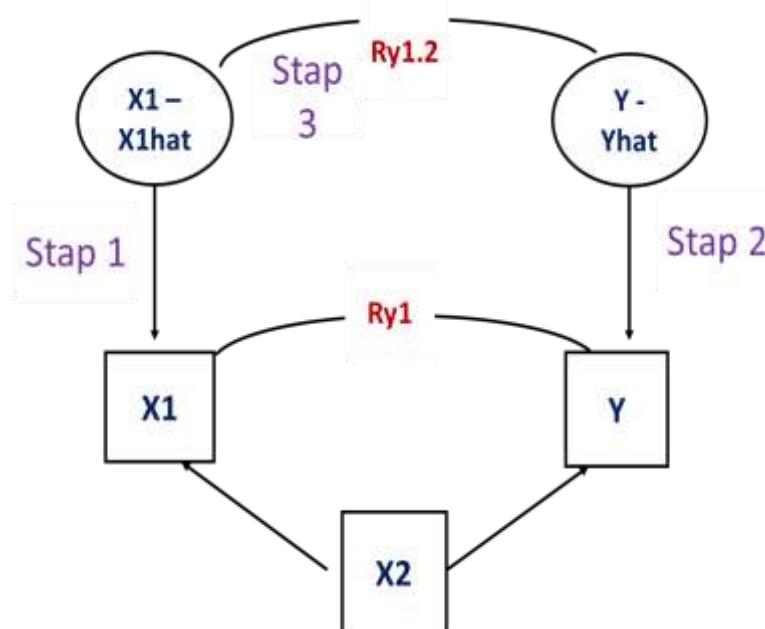
De partiële correlatie kan berekend worden aan de hand van een reeks van regressievergelijkingen.

### 3. De berekening van de partiële correlatiecoëfficiënt ahv regressievergelijkingen

Hierna volgt een uitgewerkt voorbeeld van de berekening van de partiële correlatiecoëfficiënt op basis van 3 variabelen, X1 , X2 en Y.

Figuur 3 geeft een schematische voorstelling van de werkwijze.

*Figuur 3: Schematische voorstelling van een partiële correlatie analyse op basis van 3 variabelen X1, X2 en Y*



We nemen als uitgangspunt de samenhang tussen X1 en Y (= **Ry1**) en onderzoeken de partiële correlatie tussen X1 en Y onder controle van X2 (= **Ry1.2**).

De beste manier om het principe van partiële correlatieanalyse te begrijpen is door de variantie die X2 deelt met X1 en die X2 deelt met Y te verwijderen zodat de partiële correlatie de zero-order correlatie (= zonder controlevariabelen) is tussen de residuele termen X1 – X1hat en Y – Yhat. De berekening verloopt in drie stappen.

### **STAP 1 : Verwijder de variantie die X1 en X2 met elkaar delen**

- Voer een bivariate regressie analyse uit van X1 op X2
- Bereken de verwachte waarden voor X1 (=X1hat) op basis van de regressievergelijking  
$$X1hat = a + b.X2$$
- Bereken de residuele waarden: X1 - X1hat

### **STAP 2 : Verwijder de variantie die Y en X2 met elkaar delen**

- Voer een bivariate regressie analyse uit van Y op X2
- Bereken de verwachte waarden voor Y (= Yhat) op basis van de regressievergelijking :  
$$Yhat = a + b.X2$$
- Bereken de residuele waarden: Y – Yhat

### **STAP 3 : bereken de partiële correlatiecoëfficiënt: ry1.2**

- Bereken de correlatie tussen de residuele termen : X1 – X1hat en Y – Yhat.
- De zero-order correlatie tussen de residuele termen is de partiële correlatiecoëfficiënt **ry1.2** (= notatie voor de correlatie tussen Y en X1 onder controle van X2).

### **STAP 4 : voer een significantietoets uit van de partiële correlatiecoëfficiënt**

- **bereken de t-ratio**

$$R_{xy.z} * \sqrt{N-3}$$

$$t = \frac{R_{xy.z} * \sqrt{N-3}}{\sqrt{1 - R^2_{xy.z}}}$$

- **zoek de kritieke t-waarde op in de t-tabel, gegeven het aantal vrijheidsgraden en een gekozen  $\alpha$  (=kans op een type I-fout)**
-

*Data-matrix: een fictief voorbeeld*

De statistische analyse-eenheden zijn individuen ( $N=10$ ). Alle variabelen zijn gemeten op metrisch meetniveau met scores die gaan van 0 tot 9.

- $Y$  = afhankelijke variabele
- $X_1$  = onafhankelijke variabele
- $X_2$  = controle variabele.

	<b>Y</b>	<b>X1</b>	<b>X2</b>
	3	2	1
	3	1	2
	2	5	3
	6	2	3
	4	3	4
	5	6	5
	6	5	6
	4	8	7
	9	7	8
	8	3	9
gemiddelde	<b>5,0</b>	<b>4,2</b>	<b>4,8</b>
standaardafwijking	<b>2,261</b>	<b>2,348</b>	<b>2,658</b>

<b>Correlatiematrix</b>	<b>Y</b>	<b>X1</b>	<b>X2</b>
<b>Y</b>	1		
<b>1</b>	0,251	1	
<b>X2</b>	0,777	0,612	1

**STAP 1 : Verwijder de variantie die X1 en X2 met elkaar delen**

(voer een regressie analyse uit van X1 op X2)

$$\mathbf{X1hat = a + b.X2}$$

$$\mathbf{X1hat = 1,60 + 0,54.X2}$$

X1	X2	X1hat (=voorspelde waarden X1 op basis van X2)	X1-X1hat (= residu)
2	1	2,14 (= 1.60 + 0.54.1)	-0,14 (= 2 - 2.14)
1	2	2,69 (= 1.60 + 0.54.2)	-1,69 (= 1 - 2.69)
5	3	3,23	1,77
2	3	3,23	-1,23
3	4	3,77	-0,77
6	5	4,31	1,69
5	6	4,85	0,15
8	7	5,39	2,61
7	8	5,93	1,07
3	9	6,47	-3,47

**STAP 2 : Verwijder de variantie die Y en X2 met elkaar delen**

(voer een regressie analyse uit van Y op X2)

$$\hat{Y} = a + b \cdot X_2$$

$$\hat{Y} = 1,83 + 0,66 \cdot X_2$$

Y	X2	Yhat (=voorspelde waarden Y op basis van X2)	Y - Yhat (=residu)
3	1	2,49 (= 1.83 + 0.66.1)	0,51 (= 3 - 2,49)
3	2	3,15	-0,15
2	3	3,81	-1,81
6	3	3,81	2,19
4	4	4,47	-0,47
5	5	5,13	-0,13
6	6	5,79	0,21
4	7	6,45	-2,45
9	8	7,11	1,89
8	9	7,77	0,23

**STAP 3 : Bereken de partiële correlatiecoëfficiënt: ry1.2**

X1 - X1hat	(X1 - X1hat) <sup>2</sup>	Y - Yhat	(Y - Yhat) <sup>2</sup>	(X1 - X1hat)(Y - Yhat)
-0.14	0.02	0.51	0.26	-0.07
-1.69	2.86	-0.15	0.02	0.25
1.77	3.13	-1.81	3.28	-3.20
-1.23	1.51	2.19	4.80	-2.69
-0.77	0.59	-0.47	0.22	0.36
1.69	2.86	-0.13	0.02	-0.22
0.15	0.02	0.21	0.04	0.03

2.61	6.81	-2.45	6.00	-6.39
1.07	1.14	1.89	3.57	2.02
-3.47	12.04	0.23	0.05	-0.80
	30.98		18.26	-10.71

$$Ry1.2 = \frac{\sum(X1 - X1\hat{)}(Y - Y\hat{)}}{\sqrt{\sum(X1 - X1\hat{)}^2 \sum(Y - Y\hat{)}^2}} = \frac{-10.71}{\sqrt{30.98 * 18.26}} = \frac{-10.71}{23.78} = -0.45$$

Een partiële correlatiecoëfficiënt kan net zoals de correlatiecoëfficiënt een waarde aannemen die varieert van -1 tot +1. De interpretatie ervan is identiek aan de correlatiecoëfficiënt.

Hier: de partiële correlatiecoëfficiënt tussen Y en X1 onder controle van X2 bedraagt -0.45 : dit is een matig-sterke negatieve samenhang (hoe meer...., hoe minder....).

#### STAP 4 : voer een significantietoets uit van de partiële correlatiecoëfficiënt

De statistische significantie van een partiële correlatie kan worden getest aan de hand van een t-ratio. We gebruiken volgende formule (toegepast op het voorbeeld) :

$$Ry1.2 * \sqrt{N-3}$$

$$t = \frac{Ry1.2 * \sqrt{N-3}}{\sqrt{1 - R^2 y1.2}}$$

Waarbij       $t = t$  ratio als statistische significantietoets voor de partiële correlatiecoëfficiënt

$N$  = aantal paren scores voor X1 en Y (hier = 10)

$Ry1.2$  = waarde van de partiële correlatiecoëfficiënt (hier = -.45)

$$t = \frac{-0.45 * \sqrt{10-3}}{\sqrt{1 - 0.20}} = \frac{-1.19}{0.89} = -1.34$$

$$Df = N - 3 = 7$$

$$\alpha = .05$$

kritieke t-waarde in t-tabel : 2.365

Aangezien de t-ratio de kritieke t-waarde gevonden in de t-tabel, voor 7 vrijheidsgraden en een  $\alpha = .05$ , niet overschrijdt, besluiten we dat de partiële correlatiecoëfficiënt van -.45 voor 10 statistische eenheden niet statistisch significant is. We kunnen het resultaat niet veralgemenen naar de populatie omdat we niet kunnen uitsluiten dat het gevonden resultaat in deze steekproef ‘toevallig’ is (uiteraard hebben we hier te maken met een zeer kleine steekproef: N=10).

### **Besluit:**

Het resultaat van de partiële correlatie analyse is in dit voorbeeld nogal drastisch: de samenhang tussen X1 en Y (= Ry1) bedraagt aanvankelijk  $r=0.25$ . Dit is een zwakke positieve relatie. Wanneer we echter controleren voor X2 wordt deze zwakke positieve relatie echter matig-sterk en negatief.

De partiële correlatiecoëfficiënt **Ry1.2 = -0.45**

In veel gevallen zal het zo zijn dat het bivariate verband tussen 2 variabelen zwakker wordt of zelfs volledig verdwijnt wanneer gecontroleerd wordt voor een derde variabele. Een partiële correlatie mag echter niet naïef begrepen worden als het verdwijnen van een samenhang tussen twee variabelen onder controle van een derde variabele. Ee partiële correlatie kan drastische gevolgen hebben, zoals we zagen in voorgaand voorbeeld. Waar de oorspronkelijke samenhang tussen X1 en Y zwak positief is (hoe meer X1, hoe meer Y), wordt deze niet alleen sterker onder controle van X2, maar verandert deze ook van teken (de samenhang wordt negatief).

**Concreet betekent dit dat de relatie tussen X1 en Y voor personen met gelijke scores op X2 matig-sterk èn negatief wordt (hoe meer X1, hoe minder Y)!**

De samenhang tussen X1 en Y is daarom een misleidende weergave van de werkelijkheid omdat de variantie die beide variabelen met elkaar delen hoofdzakelijk het resultaat is van de variantie die zij delen met X2.

Tabel 1 geeft een overzicht van mogelijke wijzigingen in de initiële samenhang tussen twee variabelen als gevolg van een partiële correlatie analyse en wat je hieruit als onderzoeker kan besluiten.

Tabel 1 : Mogelijke wijzigingen in bivariate samenhang als gevolg van partiële correlatie analyse

Initiële samenhang tussen X en Y	Correlatie tussen X en Y onder controle van Z	Omschrijving	Besluit
R <sub>XY</sub> = 0.50	R <sub>XYZ</sub> = 0.50	De initiële samenhang tussen X en Y onder controle van Z verandert niet.	<i>Een partiële correlatie maakt geen enkel verschil in de samenhang tussen X en Y.</i> <i>Z is noch een gemeenschappelijke oorzaak, noch een intermediaire variabele in de relatie tussen X en Y.</i>
R <sub>XY</sub> = 0.50	R <sub>XYZ</sub> = 0.00	De initiële samenhang tussen X en Y onder controle van Z verdwijnt volledig.	<i>Z is mogelijk een gemeenschappelijke oorzaak (= antecedent variabele) van de initiële samenhang tussen X en Y (spurieuze relatie).</i> <i>Z kan ook een intermediaire variabele zijn in de relatie tussen X en Y (indirecte relatie).<sup>20</sup></i>
R <sub>XY</sub> = 0.05	R <sub>XYZ</sub> = 0.60	De initiële samenhang tussen X en Y onder controle van Z verandert van een zeer lage samenhang naar een sterke samenhang.	<i>Z is een suppressor-variabele in de relatie tussen X en Y. Z 'verbergt' de relatie tussen X en Y, die zichtbaar wordt wanneer de gedeelde variatie tussen X en Z en Y en Z 'verwijderd' wordt.</i>

<sup>20</sup> Het feit dat Z een gemeenschappelijke oorzaak (of antecedent variabele) is dan wel een intermediaire variabele die tussen de relatie van X en Y in staat, wordt bepaald op niet-statistische gronden. Op basis van een statistische partiële correlatie analyse kan enkel beslist worden of de samenhang tussen X en Y al dan niet wijzigt (behouden blijft, verzwakt, verdwijnt of versterkt).

#### 4. Berekening van de partiële correlatiecoëfficiënt ahv rekenkundige formule

In de vorige paragraaf hebben we de partiële correlatiecoëfficiënt berekend aan de hand van een reeks regressievergelijkingen. Er bestaat echter een rekenkundige formule om uit de bivariate correlaties de partiële correlatie te berekenen.

$$\frac{r_{X_1Y} - (r_{X_1X_2})(r_{YX_2})}{\sqrt{1 - r_{X_1X_2}^2} \times \sqrt{1 - r_{YX_2}^2}}$$

We geven een voorbeeld:

$$\begin{aligned} X \text{ versus } Y: r_{XY} &= +.50 \quad r_{XY}^2 = .25 \\ X \text{ versus } Z: r_{XZ} &= +.50 \quad r_{XZ}^2 = .25 \\ Y \text{ versus } Z: r_{YZ} &= +.50 \quad r_{YZ}^2 = .25 \end{aligned}$$

$$\begin{aligned} r_{XY \cdot Z} &= \frac{0.50 - (0.50)(0.50)}{\sqrt{1 - 0.25} \times \sqrt{1 - 0.25}} \\ &= +0.33 \\ \text{Besluit } r_{XY \cdot Z}^2 &= 0.11 \end{aligned}$$

Dezelfde redenering kan worden toegepast om de partiële correlatie te berekenen tussen X en Z, waarbij we de gemeenschappelijke variatie met Y moeten verwijderen.

$$\begin{aligned} r_{XZ} - (r_{XY})(r_{YZ}) \\ r_{XZ \cdot Y} = \frac{\sqrt{1 - r_{XY}^2} \times \sqrt{1 - r_{YZ}^2}}{\sqrt{1 - r_{XY}^2} \times \sqrt{1 - r_{YZ}^2}} \end{aligned}$$

en voor de berekening van de partiële correlatie tussen Y en Z, waarbij de effecten van X verwijderd zijn:

$$r_{YZ} = \frac{(r_{XY})(r_{XZ})}{\sqrt{1 - r_{XY}^2} \times \sqrt{1 - r_{XZ}^2}}$$

$$r_{YZ \cdot X} = \frac{(r_{YZ} - (r_{XY})(r_{XZ}))}{\sqrt{1 - r_{XY}^2} \times \sqrt{1 - r_{XZ}^2}}$$

Laten we een rekenvoorbeeld geven uit een criminologisch onderzoek. De variabelen C (crime / delinquency), A (Association delinquent peers) en V (low self-control). Er is duidelijk overlap tussen de drie theoretische variabelen uit de criminale etiologie. Er is een positief verband tussen lage zelfcontrole en delinquentie, tussen delinquenten en lage zelfcontrole en tussen delinquentie en delinquenten.

<b>C versus A:</b>	$r_{CA} = +0.49$	$r^2_{CA} = 0.24$
<b>C versus V:</b>	$r_{CV} = +0.73$	$r^2_{CV} = 0.53$
<b>A versus V:</b>	$r_{AV} = +0.59$	$r^2_{AV} = .035$

$$r_{CA} = \frac{(r_{CV})(r_{AV})}{\sqrt{1 - r_{CV}^2} \times \sqrt{1 - r_{AV}^2}}$$

$$r_{CA \cdot V} = \frac{r_{CA} - (r_{CV})(r_{AV})}{\sqrt{1 - r_{CV}^2} \times \sqrt{1 - r_{AV}^2}}$$

$$= \frac{0.49 - (0.73)(0.59)}{\sqrt{1 - 0.73^2} \times \sqrt{1 - 0.59^2}}$$

$$r_{CA \cdot V} = \frac{r_{CA} - (r_{CV})(r_{AV})}{\sqrt{1 - r_{CV}^2} \times \sqrt{1 - r_{AV}^2}}$$

$$r_{CA \cdot V} = +.11$$

$$\text{Besluit } r^2_{CA \cdot V} = 0.01$$

Hier zien we dat het bivariate verband een drastische overschatting van de realiteit laat zien. Wanneer we controleren voor V, dan is het verband tussen C en A quasi weg. Immers, een correlatie van 0.11 is niet echt noemenswaardig te noemen. Effecten onder .2 beschouwen we als uiterst zwak.

## 5. Suppressie-effect

Laten we tot slot het **suppressie-effect** nog even onder de loep nemen. Dit is belangrijk want het komt in de praktijk van het onderzoek wel eens voor en beginnende onderzoekers zijn zich er niet altijd van bewust. Dat heeft tot gevolg dat men resultaten verkeerd inschat en dat men een grotere kans heeft om nonsens-beleid te voeren. We hernemen het voorbeeld van hierboven. Veronderstel dat een professor metingen heeft uitgevoerd op drie kenmerken: inspanningen van studenten om een vak in te studeren, de scores van de studenten en de angst voor een vak bij studenten. Die drie kenmerken hebben elk een associatie met elkaar.

X = Inspanningen gedaan voor het studeren van het vak

Y = De score van de student

Z = De mate waarin de student angst heeft voor het vak

Hier zijn de bivariate correlaties tussen de drie variabelen:

X versus Y:  $r_{XY} = +0.20$     $r^2_{XY} = 0.04$

X versus Z:  $r_{XZ} = +0.80$     $r^2_{XZ} = 0.64$

Y versus Z:  $r_{YZ} = -0.40$     $r^2_{YZ} = 0.16$

Is het niet vreemd dat het verband tussen de scores van de student en de inspanningen gedaan door de student zo laag is? ( $r_{XY}=+.20$  en  $r^2_{XY}=.04$ ). Als je nader onderzoek doet, zal je zien wat er werkelijk aan de hand is. Als de angst voor het vak groter wordt, dan studeren studenten meer voor het vak, vandaar  $r_{XZ}=+.80$  en  $r^2_{XZ}=.64$ . Aan de andere kant is het ook zo dat angst voor een vak en de score voor het vak negatief met elkaar correleren: angst kan de concentratie helemaal verstören ( $r_{YZ}=-.40$  en  $r^2_{YZ}=.16$ ). Laten we het onderdrukkende effect van de angst voor het vak verwijderen en de bivariate correlatie tussen scores en inspanningen opnieuw bekijken. Wat zien we na wat rekenwerk?

$$0.20 - (0.80)(-0.40)$$

$$r_{XY \cdot Z} = \frac{0.20 - (0.80)(-0.40)}{\sqrt{1 - 0.64} \times \sqrt{1 - 0.16}}$$

$$r_{XY \cdot Z} = +0.95$$

$$\boxed{r_{XY} = 0.20 \quad r^2_{XY} = 0.04 \\ r_{XZ} = 0.80 \quad r^2_{XZ} = 0.64 \\ r_{YZ} = -0.40 \quad r^2_{YZ} = 0.16}$$

De bivariate correlatie die eerder zwak was, heeft plaats gemaakt voor een heel sterke partiële correlatie: van een kleine 0.20 (bivariaat) naar een impressionante +0.95 (partieel).

## **6. Leerdoelen**

Studenten dienen de grenzen van de bivariate statistiek te kennen en zelf in staat te zijn een partiële correlatie tussen twee variabelen onder de statistische controle van een derde variabele te berekenen. In werkelijkheid is het denkbaar dat de partiële correlatie wordt berekend tussen twee variabelen onder de controle voor meerdere variabelen, echter deze extensie maakt niet het voorwerp uit van dit verkennend handboek. Studenten dienen te begrijpen waarom multivariate analyses zo belangrijk zijn in de criminologie. Het principe van de statistische controle moet goed begrepen worden.

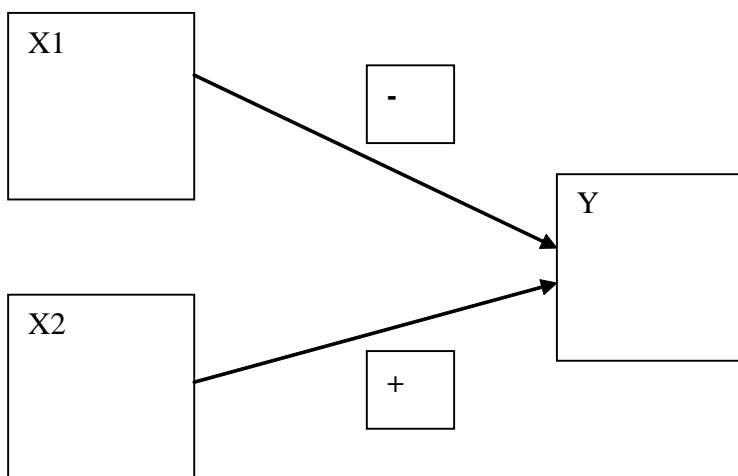
## Hoofdstuk 11

### Regressieanalyse met twee onafhankelijke variabelen

#### 1. Inleiding

De **meervoudige (of multipele) regressieanalyse** wordt gebruikt wanneer men geïnteresseerd is in het verklaren van de spreiding in een afhankelijke variabele van het metrische niveau op basis van meerdere onafhankelijke variabelen die eveneens van het metrische niveau zijn. Bovendien: de werkelijkheid is complex en complexe fenomenen kunnen niet zomaar door één en slechts één kenmerk verklaard worden. De multipele regressie laat toe om de rechtstreekse statistische effecten van meerdere onafhankelijke variabelen tegelijkertijd te bestuderen. Dit basisformat (het pijltjesdiagram) heeft wat men noemt een **convergente causale structuur**, dit wil zeggen: men is geïnteresseerd in de **rechtstreekse effecten van meerdere onafhankelijke variabelen (predictoren) op één afhankelijke variabele (het explanandum)**.

**Figuur: causaal diagram van een meervoudige regressie**



We geven een voorbeeld aan de hand van twee onafhankelijke variabelen, X1 en X2.

- X1= de mate van sociale controle van de ouders op de adolescent. We noemen deze variabele verderop “sociale controle ouders”.
- X2= de mate waarin de adolescent het overtreden van regels als moreel aanvaardbaar beschouwt: We noemen deze variabele verderop “criminele waarden van adolescent”.

- $Y =$  de frequentie van regelovertredend gedrag van de adolescent.

Vooraleer we als criminoloog-onderzoeker de analyse uitvoeren, is het belangrijk de veronderstellingen die de onderzoeker heeft, neer te schrijven. Waarom? Dit laat toe om achteraf te controleren of de veronderstellingen juist waren of fout. Veronderstellingen zijn vaak afgeleid uit theorie en het is belangrijk de terugkoppeling te maken.

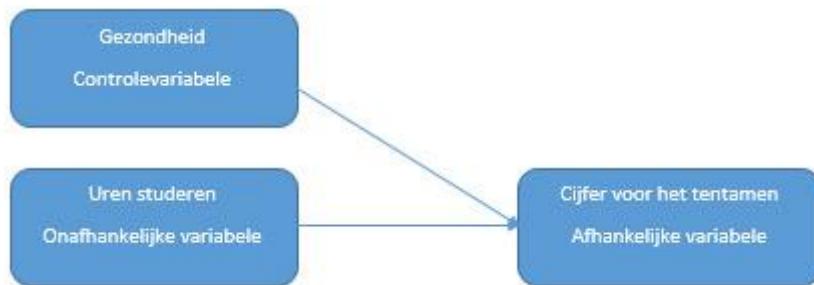
**Hypothese 1:** de mate van sociale controle van de adolescent heeft een **remmend** (= negatief) effect op de frequentie van regelovertrding, *onafhankelijk van de mate waarin de adolescent het overtreden van regels als moreel aanvaardbaar beschouwt.*

**Hypothese 2:** De mate waarin de adolescent criminale waarden als moreel aanvaardbaar beschouwt, heeft een **duwend** (= positief) effect op de frequentie van regelovertrding van de adolescent, *onafhankelijk van de sociale controle van de ouders*

## 2. De noodzaak voor het meten van controlevariabelen

Een controlevariabele (**confounder**) is een variabele die je meeneemt in je onderzoek, maar waar niet speciaal je aandacht naar uitgaat. Je neemt de variabele wel mee omdat deze invloed heeft op de afhankelijke variabele en omdat deze variabele ook samenhangt met de onafhankelijke variabele. De controlevariabelen weglaten uit het onderzoek zou betekenen dat de resultaten van je onderzoek minder accuraat zijn. Dit is vooral aan de orde wanneer je een statistische analyse doet en je een bepaalde oorzaak-gevolgrelatie statistisch wilt bewijzen. Sommige controlevariabelen kunnen geen oorzaken zijn, maar toch worden ze mee opgenomen. De reden waarom men er zoveel aandacht aan besteedt, is omdat men vreest dat een causale relatie zou verdwijnen onder de controle van de controlevariabele. Als men kan zeggen dat de resultaten van een studie, waarbij de relatie tussen enkele causale factoren en regelovertredend gedrag wordt bestudeerd, blijft bestaan onder controle van een aantal potentieel verstorende factoren, dan komt dat het geloof dat men zal hechten aan de resultaten ten goede. Wanneer we bijvoorbeeld het verband onderzoeken tussen de variabelen uit één theorie, en we kunnen zeggen dat het verband stand houdt onder controle van de variabelen uit een concurrerende theorie, dan is dit sterk.

## **Controlevariabelen in een conceptueel model**



## **Voorbeeld controlevariabele in conceptueel model**

In deze paragraaf geven we het voorbeeld van de student die veel studeert (onafhankelijke variabele) om zo een hoger cijfer te halen voor zijn tentamen (afhangelijke variabele). We voegen nu een controlevariabele aan het voorbeeld toe. Een voorbeeld van een controlevariabele in ons voorbeeld is ‘Gezondheid’. Het valt te argumenteren dat een slechte gezondheid ervoor zorgt dat de student een lager cijfer haalt voor het tentamen (invloed op afhangelijke variabele). Wanneer je kwantitatief onderzoekt onder studenten om statistisch aan te tonen dat er een statistisch effect is van het aantal ‘Uren studeren’ en het behalen van ‘Cijfer voor het tentamen’ dan is het goed om additioneel te vragen naar de gezondheid van de student. Doe je dit niet, en neem je de variabele ‘Gezondheid’ dus ook niet mee als controlevariabele in je onderzoek, dan kan dit betekenen dat het aantonen van de oorzaak-gevolgrelatie een stuk moeilijker is. Het verband dat je vindt, kan namelijk voor een deel te wijten zijn aan de gezondheid van de student. Er zijn theoretisch verschillende scenario’s mogelijk: gezondheid beïnvloedt hoe lang je kan studeren en heeft zo een indirect effect op je punten op je tentamen. Maar het kan zijn dat het rechtstreeks effect van het aantal uren studeren op het tentamencijfer verdwijnt wanneer je ook rekening houdt met gezondheid, en dat is een probleem, want dan vervalt het empirisch bewijs voor de hypothese dat er een rechtstreeks verband bestaat tussen het aantal uren studeren en het cijfer voor het tentamen. We willen de lezer er op wijzen dat controlevariabelen gebruiken belangrijk is, maar nooit ondoordacht mag gebeuren. Er moet een theoretische reden achter schuilgaan. Zo kan men zich de vraag stellen of we moeten controleren voor variabelen die geen oorzaken kunnen zijn in de betekenis die we aan oorzaakelijkheid geven (productie, aanwezigheid van een mechanisme, manipuleerbaarheid). Echter, de traditie van het gebruiken van

controlevariabelen is zo sterk ingebakken in de sociale wetenschappen, dat men op de duur vergeet waarom men bepaalde controlevariabelen toevoegt aan de analyse.

### **3. De vergelijking tussen twee bivariate versus één meervoudige regressie**

We spraken reeds over de mogelijke invloed van storende variabelen die samenhangen met de onafhankelijke variabele en die ook invloed uitoefenen op de afhankelijke variabelen. De onderlinge samenhang tussen onafhankelijke variabelen is op zich een realiteit die maakt dat we niet zomaar verschillende bivariate regressieanalyses kunnen uitvoeren als vervanging voor een meervoudige regressieanalyse. *De samenhang tussen onafhankelijke variabelen noemen we met een statistisch begrip “multicollineariteit”*. In een lineaire regressieanalyse vertaalt zich deze samenhang in een correlatie tussen de onafhankelijke variabelen X1 en X2. Hoe sterker twee onafhankelijke variabelen samenhangen, hoe meer misleidend de uitkomsten van een bivariate analyse kunnen zijn. De samenhang tussen twee onafhankelijke variabelen mag ook niet te hoog zijn als we een meervoudige regressieanalyse willen uitvoeren. *Is deze samenhang hoger dan .50 dan dient men voorzichtig te zijn bij de interpretatie van de resultaten en is deze .80 of hoger, dan mag men absoluut geen meervoudige regressie met die onafhankelijke variabelen uitvoeren die dergelijke hoge samenhang hebben. De beide variabelen kunnen dan analytisch nauwelijks van elkaar onderscheiden worden en de resultaten van dergelijke analyse zijn hoogst onbetrouwbaar. De gestandaardiseerde effectparameters (richtingscoëfficiënten) kunnen dan onmogelijke waarden bekomen (groter dan 1 of kleiner dan -1).*

Daarom dient men bij een multivariate analyse steeds de voorbereidende controle op multicollineariteit door te voeren. Stel dat we geïnteresseerd zijn in het causale effect van criminale waarden en normen en de sociale controle in het gezin bij het verklaren van individuele verschillen in de frequentie van regelovertredend gedrag van jonge adolescenten. Waarom is het niet zonder gevaren om twee afzonderlijke bivariate regressieanalyses uit te voeren en de determinatiecoëfficiënten van beide afzonderlijke analyses bij elkaar op te tellen? Dit zou toch ogenschijnlijk het meest voor de hand liggen?

De partiële overlap of onderlinge samenhang tussen de beide onafhankelijke variabelen is de reden waarom we geen twee afzonderlijke bivariate analyses met elkaar mogen optellen. In het voorbeeld dat overigens kan worden overgedaan aan de hand van de oefendatabestanden, is de overlap tussen de twee onafhankelijke variabelen X1 en X2 reëel en bedraagt -0.455.

**Correlations**

		sociale controle ouders	criminele waarden
sociale controle ouders	Pearson Correlation	1	-,455**
	Sig. (2-tailed)		,000
	N	1521	1514
criminele waarden	Pearson Correlation	-,455**	1
	Sig. (2-tailed)	,000	
	N	1514	1542

\*\*. Correlation is significant at the 0.01 level (2-tailed).

Laten we nu eens de uitkomsten van de afzonderlijke bivariate regressieanalyses bekijken en vergelijken met de uitkomsten van één meervoudige regressieanalyse. We presenteren de gestandaardiseerde regressiecoëfficiënten omdat deze toelaten de vergelijking te maken tussen de sterkte van het effect van elke variabele op de afhankelijke variabele.

**Tabel: Twee afzonderlijke bivariate regressieanalyses en één meervoudige regressieanalyse**

Afhankelijke variabele: regeloverredend gedrag	Model 1(bivariaat) $\beta$	Model 2 (bivariaat) $\beta$	Model 3 (multivariaat) $\beta$
Sociale controle ouders (X1)	-0.382*	--	-0.198*
Criminele waarden (X2)	--	0.496*	0.404*
R-kwadraat	<b>14.6%</b>	<b>24.6%</b>	<b>27.5%</b>

\* p < 0.05 of beter

Beide onafhankelijke variabelen verklaren een substantieel deel van de variantie in crimineel gedrag van jongeren: X1 verklaart 14.6% en X2 verklaart 24.6%. We kunnen hieruit echter niet besluiten dat beide variabelen samen 14.6% + 24.6% van de variatie in Y verklaren. Dit komt omdat de beide variabelen sterk samenhangen. Het kan niet anders of er is dus *gedeelde variantie* tussen X1 en X2, waardoor de uitvoering van twee bivariate analyses misleidend zou geweest zijn.

Bovendien leren we ook hier nog iets anders: de gestandaardiseerde richtingscoëfficiënten (de beta-waarden) zijn in de meervoudige regressieanalyse verschillend van de richtingscoëfficiënten afkomstig uit de bivariate regressieanalyses. Het effect van X1 op Y is kleiner onder controle van X2. Dit leiden we af uit de tabel. We zien dat het rechtstreeks effect van ouderlijke controle op delinquent gedrag groter is in een bivariate analyse dan in de meervoudige regressieanalyse waar we simultaan rekening houden met de invloed van criminale waarden. Het rechtstreeks effect van delinquente waarden is ook iets kleiner wanneer we rekening houden met het effect van ouderlijke controle.

#### **4. De uitbreiding naar een meervoudige regressieanalyse**

De bivariate regressievergelijking ziet er als volgt uit:

$$y = a + bx + e$$

$$\hat{y} = a + bx$$

Y is de geobserveerde score op het explanandum, a is het intercept en b is de richtingscoëfficiënt. Tenslotte is e de residuele term.

De multipele regressievergelijking is een uitbreiding van het bivariate model en ziet er als volgt uit:

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n + e$$

$$\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Het vinden van de waarden voor de verschillende *b*-waarden vergt veel rekenwerk in de situaties waarin het aantal onafhankelijke variabelen meer dan twee bedraagt ( $k > 2$ ). Een snelle uitrekening van de netto-effecten van de *b*-waarden in de situatie van drie of meerdere onafhankelijke variabelen gebeurt daarom aan de hand van de matrix algebra. Studenten kunnen gerust zijn: de matrix algebra blijft buiten het bestek van dit handboek. Om inhoudelijk te begrijpen wat er gebeurt in de situatie van twee onafhankelijke variabelen, is echter minder rekenwerk nodig dan men op het eerste zicht zou denken.

## 5. Het relatieve belang van elke onafhankelijke variabele

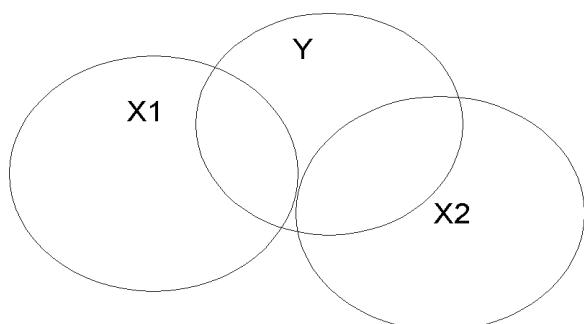
In een eenvoudige regressieanalyse met één afhankelijke variabele is er een afhankelijke variabele en een onafhankelijke variabele. De impact van de onafhankelijke variabele (hoe belangrijk is de onafhankelijke variabele bij de predictie van de afhankelijke variabele) wordt uitgedrukt aan de hand van de verklaarde variantie  $r^2$ . Als  $r^2$  1.0 bedraagt, weten we dat de afhankelijke variabele perfect kan voorspeld worden op basis van de onafhankelijke variabele. Als  $r^2$  nul bedraagt, weten we dat er alvast geen lineair verband bestaat tussen de onafhankelijke variabele en de afhankelijke variabele. In de situatie met twee of meerdere onafhankelijke variabelen verkrijgen we ook een verklaarde variantie, de totale  $R^2$ . Deze  $R^2$  vertelt ons hoeveel van de variantie in  $Y$  kan verklaard worden op basis van de onafhankelijke variabelen samen. Deze totale verklaarde variantie vertelt ons iets over de relatieve belangrijkheid van de lineaire combinatie van de onafhankelijke variabelen ( $b_1X_1+b_2X_2+\dots+b_kX_k$ ). Heel vaak willen we weten welke impact elke variabele op zichzelf heeft, onder controle van de overige onafhankelijke variabelen.

Opnieuw maken we gebruik van venndiagrammen om uit te leggen wat er gebeurt. **Venndiagrammen mogen misschien niet helemaal de meest correcte manier zijn om de idee achter een multivariate regressieanalyse uit te leggen, maar deze visuele voorstelling zorgt voor een beter begrip van het concept.**

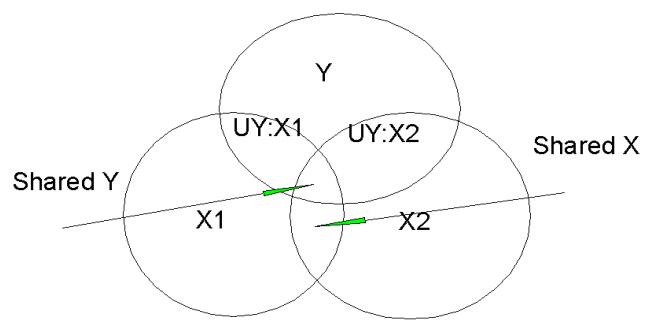
We voorspellen  $Y$  op basis van twee onafhankelijke variabelen,  $X_1$  en  $X_2$ . In de ideale situatie zijn  $X_1$  en  $X_2$  niet met elkaar gecorreleerd en dan krijgen we de situatie zoals deze is getekend in het linkse venndiagram.

**Figuur: de relatie tussen de onafhankelijke variabelen**

**Geen multicollineariteit**



**Wel multicollineariteit**



De drie cirkels stellen onafhankelijke variabelen voor. Elke cirkel represeneert de variabiliteit in de variabele. De grootte van de overlap kan worden afgerezen uit de grootte van de overlappende gebieden. In de situatie die links werd getekend, is er geen overlap tussen X<sub>1</sub> en X<sub>2</sub>, beiden hebben wel elk een impact op Y. In de situatie rechts is er wel overlap tussen X<sub>1</sub> en X<sub>2</sub> en hier stelt zich hetzelfde probleem als we eerder beschreven bij de uiteenzetting over de partiële correlatie. We presenteren de correlatiematrix tussen de drie variabelen:

**Tabel: correlatiematrix tussen de drie variabelen**

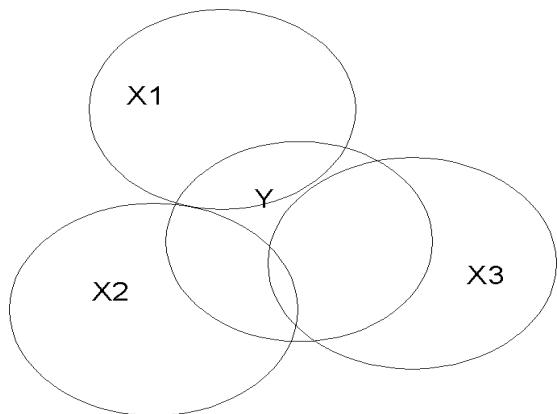
R	Y	X <sub>1</sub>	X <sub>2</sub>
Y	1		
X <sub>1</sub>	0.50	1	
X <sub>2</sub>	0.60	0.00	1

In het geval dat X<sub>1</sub> en X<sub>2</sub> niet gecorreleerd zijn kunnen we de totale verklaarde variantie op basis van X<sub>1</sub> en X<sub>2</sub> schatten door de afzonderlijke verklaarde varianties op te tellen: in ons voorbeeld is dat  $0.50^2+0.60^2 = 0.25+0.36 = 0.61$ .

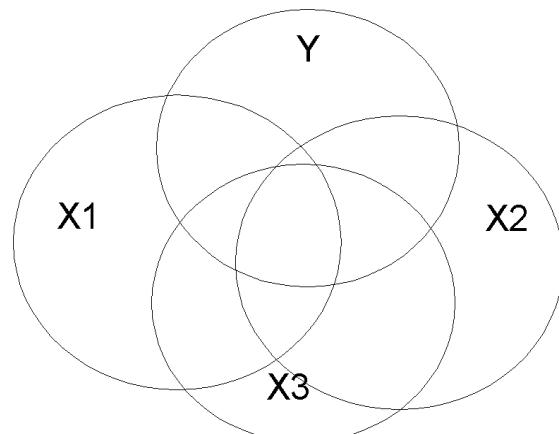
In de meeste gevallen is er echter wel een zekere mate van overlap te verwachten. In de sociale werkelijkheid hangen vele kenmerken nu eenmaal samen

Wanneer we dit idee doortrekken naar de multiplele regressie met drie onafhankelijke variabelen dan is dit ook zo:

De quasi ideale situatie



De realiteit



Linksboven zien we een quasi ideale situatie: minimale overlap tussen de verklarende variabelen. Rechtsboven zien we de realiteit: een kluwen van correlerende variabelen. Laat dit de lezer niet afschrikken.

## 6. De berekening van de gestandaardiseerde gewichten ( $\beta_1$ en $\beta_2$ )

De parameters van een multivariate regressieanalyse kunnen gemakkelijk berekend worden op basis van de correlatiematrix. Dit is goed nieuws. De formule is eenvoudig en we verkrijgen onmiddellijk de gestandaardiseerde richtingscoëfficiënt: dit is tenslotte de coëfficiënt die we het meest nodig hebben om zinvolle uitspraken te kunnen doen over het relatieve belang van elk van de onafhankelijke variabelen. Het probleem met de ongestandaardiseerde regressiegewichten is dat zij elk in hun eigen metrische eenheid worden gepresenteerd en dus elke zinvolle vergelijking in de weg staan. De betekenis van de toename in Y op basis van de toename van X1 met één eenheid wordt uitgedrukt in de meeteenheid van X1. Hetzelfde geldt voor X2.

Wanneer we de gestandaardiseerde richtingscoëfficiënten willen berekenen op basis van de correlatiematrix tussen de onafhankelijke variabelen, moeten we volgende formule gebruiken

$$\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$

en

$$\beta_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}$$

waarbij  $r_{y1}$  de correlatie van y met X1 voorstelt,  $r_{y2}$  is dan de correlatie van y met X2, en  $r_{12}$  is de correlatie tussen X1 en X2. Let er op dat de twee formules bijna identiek zijn. Het verschil zit hem in de volgorde van de symbolen in de teller.

Als onze correlatiematrix er als volgt uitziet:

	Y	X <sub>1</sub>	X <sub>2</sub>
Y	1		

X <sub>1</sub>	0.77	1	
X <sub>2</sub>	0.72	0.68	1

dan bekomen we volgende uitkomsten voor de gestandardiseerde regressiegewichten van X1 en X2:

$$\beta_1 = \frac{.77 - (.72)(.68)}{1 - .68^2} = .521577$$

$$\beta_2 = \frac{.72 - (.77)(.68)}{1 - .68^2} = .365327$$

### De berekening van de ongestandaardiseerde regressiecoëfficiënten (b1 en b2)

De berekening van de ongestandaardiseerde regressiegewichten kan eenvoudig door de gestandaardiseerde regressiegewichten te vermenigvuldigen met een coëfficiënt. Voor b1 is dat de standaardafwijking van Y gedeeld door de standaardafwijking van x1. Voor b2 is dat de standaardafwijking van y gedeeld door de standaardafwijking van x2.

$$b_1 = \left( \frac{r_{y,x1} - r_{y,x2}r_{x1,x2}}{1 - (r_{x1,x2})^2} \right) \left( \frac{SD_y}{SD_{x1}} \right)$$

$$b_2 = \left( \frac{r_{y,x2} - r_{y,x1}r_{x1,x2}}{1 - (r_{x1,x2})^2} \right) \left( \frac{SD_y}{SD_{x2}} \right)$$

### De berekening van de totale R<sup>2</sup>

In de situatie van gecorreleerde onafhankelijke variabelen bekomen we de correcte waarde voor de verklaarde variantie door de volgende formule te hanteren:

$$R^2 = \beta_1 r_{y1} + \beta_2 r_{y2}$$

Als je de logica van de regressieanalyse snapt, dan begrijp je nu ook waarom de determinatiecoëfficiënt gedoemd is om een lagere waarde te hebben dan de optelsom van de twee bivariate correlaties (in geval van overlap tussen de twee onafhankelijke variabelen). Het gaat hier, zoals de formule zegt om het relatieve aandeel van de beide onafhankelijke variabelen: in de formule wordt gebruik gemaakt van de gestandaardiseerde richtingscoëfficiënten en de partiële correlatiecoëfficiënten. Hierdoor is elke gemeenschappelijke overlap verwijderd. Het gaat hem om de optelsom van wat uniek door X1 kan verklaard worden en wat uniek door X2 kan verklaard worden.

En hoe vinden we het intercept? Het intercept kan gevonden worden op dezelfde manier als we eerder hebben aangetoond voor de situatie waarbij we slechts één onafhankelijke variabele hadden.

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

Samengevat: wat hebben we nodig om de meervoudige regressieanalyse te kunnen uitvoeren?

- Gemiddelden van alle variabelen
- Standaardafwijkingen van alle variabelen
- Correlaties tussen alle variabele

## 7. Veronderstellingen bij het uitvoeren van een lineaire regressie analyse

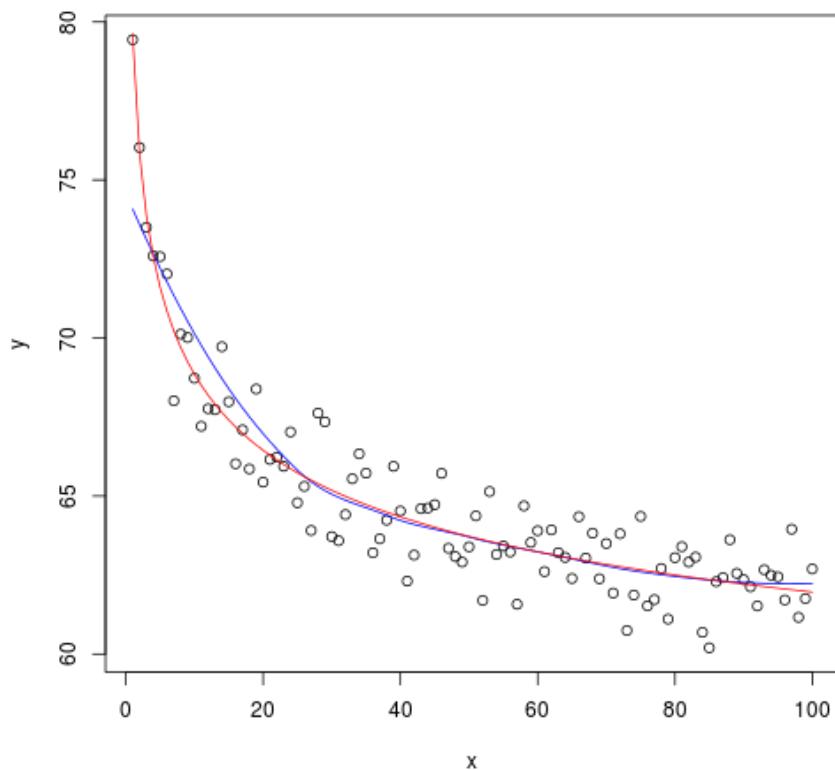
Het is belangrijk stil te staan bij een aantal niet eerder genoemde, maar toch belangrijke veronderstellingen waaraan dient voldaan te worden wanneer we een regressieanalyse uitvoeren.

**Ten eerste** wordt wat betreft het meetniveau verondersteld dat zowel de afhankelijke als de onafhankelijke variabele gemeten zijn op **interval- of ratio-meetniveau**.

**Een tweede** veronderstelling is dat er een *lineair verband* bestaat tussen de onafhankelijke en afhankelijke variabele. Of beter nog, de lineaire regressie levert maar zinvolle informatie op voor zover er sprake is van een lineair verband tussen beide variabelen. Deze veronderstelling kan nagegaan worden door een grafische voorstelling te maken waar de residuele termen

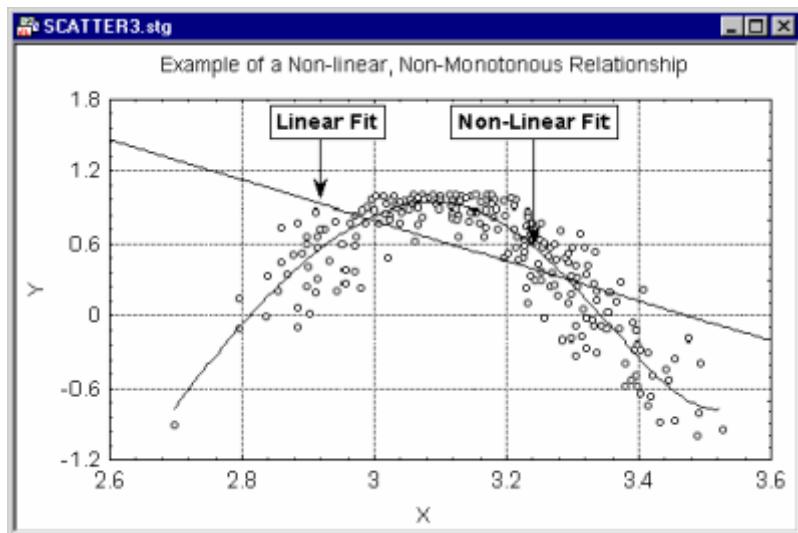
worden uitgezet tegenover de waarden van de onafhankelijke variabele. Indien deze residuen een (niet-lineair) patroon vertonen, is er sprake van schending van de assumptie van lineariteit. Ook het grafisch weergeven (“plotten”) van de afhankelijke variabele tegenover de onafhankelijke variabele kan je hierover al een idee geven.

### Voorbeeld een: non-lineariteit



In het eerste voorbeeld is non-lineariteit weliswaar aanwezig, maar het verband tussen x en y is negatief, ongeacht je nu een lineaire associatiemaaat of een non-lineaire associatiemaaat berekent.

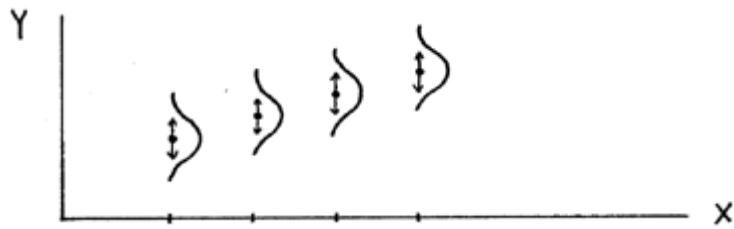
## Voorbeeld twee: non-lineariteit



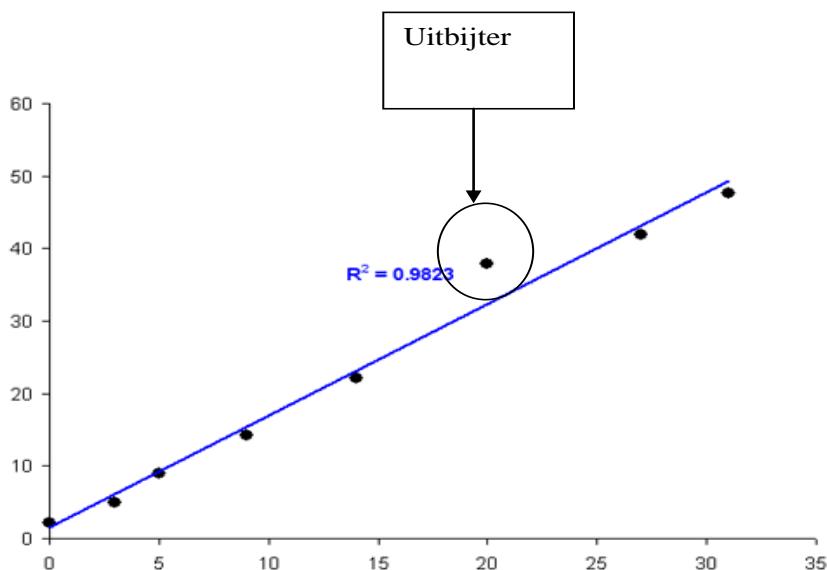
In dit voorbeeld wordt het wel erg duidelijk wat er kan gebeuren als je geen rekening houdt met de nonlineariteit. In dit voorbeeld is het zo dat een lineaire analyse een negatief verband suggereert en een non-lineaire analyse toont aan dat het verband tussen x en y eerst stijgt en dan vanaf een bepaalde waarde terug afneemt.

**Ten derde** wordt er verondersteld dat voor elke waarde van de onafhankelijke variabele, de afhankelijke variabele een **normale verdeling** kent. Bovendien is het noodzakelijk dat de variatie van de afhankelijke variabele voor elke waarde van de onafhankelijke variabele gelijk is. De technische term voor deze veronderstelling is *homoscedasticiteit*. Wanneer deze variaties verschillen naargelang de waarde van de onafhankelijke variabele spreekt men van *heteroscedasticiteit*. Mogelijke schendingen van deze assumpties kunnen eveneens nagegaan worden door het plotten van de residuen tegenover de onafhankelijke variabele. Het omgekeerde van homoscedasticiteit heet heteroscedasticiteit. Dit betekent dat de spreiding van geobserveerde waarden voor Y groter wordt naarmate dat de waarde op de onafhankelijke waarde toeneemt. Of de assumptie van homoscedasticiteit geschonden is, kan onder meer worden nagegaan door een aantal statistische toetsen. Een pionier was *White's test*.

**Figuur: Grafische illustratie van homoscedasticiteit**



**Ten vierde** dienen we ook voorzichtig te zijn met uitbijters of “**outliers**”. Dit zijn extreme waarden die de regressierechte beïnvloeden. Een puntenwolk toont je niet enkel de spreiding van observaties in het tweedimensionale vlak, maar als je de best passende lijn opvraagt, dan zie je ook welke punten sterk afwijken. We noemen deze punten uitbijters en men dient er voor waakzaam te zijn. Ze kunnen de relatie tussen twee variabelen ernstig beïnvloeden.



**Ten vijfde** mag er geen autocorrelatie zijn. **Autocorrelatie** (of **seriële correlatie**) betekent dat de waarden van een variabele gemeten op T2 niet beïnvloed worden door de waarden die een statistische eenheid had op T1. Dit probleem is eigen aan de tijdsreeksanalyse, die tot doel heeft voorspellingen te maken op basis van gegevens die doorheen de tijd werden verzameld. De analyse-eenheid is dan het jaar en de variabelen zijn bijvoorbeeld de werkloosheidsgraad en de criminaliteitsgraad. De criminaliteitsgraad van een eenheid op T1 is niet onafhankelijk

van de criminaliteitsgraad van diezelfde variabele op T2. Als er seriële correlatie is, moeten robuuste analysetechnieken worden gekozen. In de econometrie, de toepassing van de statistiek in het domein van de economie, wordt dit vaak gedaan en bestaan diverse manieren om aan tijdreeksanalyse te doen.

**Ten zesde** mag er geen statistische interactie zijn. Dit wordt later uitgelegd wanneer we het hebben over de grenzen van de bivariate statistiek. Een ander woord is conditionele causaliteit.

## 8. Controle op de regressievoorwaarden

In deze paragraaf wordt nagegaan in welke mate er een schending is van de regressievoorwaarden. De controles die zullen besproken worden, hebben betrekking op normaliteit, heteroscedasticiteit, lineariteit, en additiviteit. We zullen eveneens oog hebben voor uitbijters. Autocorrelatie laten we buiten beschouwing omdat dit probleem vooral een rol speelt in tijdreeksanalyses. Het probleem van multicollineariteit wordt niet meer afzonderlijk behandeld omdat we uit wat voorafging weten dat de onderlinge correlaties tussen de latente variabelen in geen geval problematisch zijn.

### Normaliteit

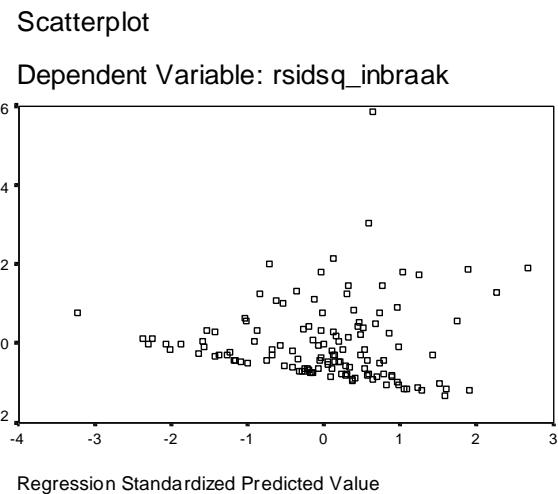
Het gaat hier zowel om univariate als multivariate normaliteit. Zowel de afhankelijke als de onafhankelijke variabelen dienen te voldoen aan deze eis.

### Heteroscedasticiteit

Homoscedasticiteit betekent dat de spreiding in de error termen constant is. Wanneer de spreiding van de error termen toeneemt met de voorspelde waarden, komen we technisch gezien in de problemen bij de interpretatie van de significantietoets. In dit geval spreken we van **heteroscedasticiteit** en wordt de standaardfout van de parameters onderschat, met als gevolg dat effecten te snel als significant worden beschouwd.

Het waarnemen van heteroscedasticiteit met het blote oog is niet altijd vanzelfsprekend. Detectie van heteroscedasticiteit is op verschillende manieren mogelijk. Eén van de mogelijkheden om het probleem van de heteroscedasticiteit te onderzoeken, is het gebruik van

**White's test.** White ontwikkelde deze methode vanuit de econometrie<sup>17</sup>. In het geval van heteroscedasticiteit is OLS-regressie niet de meest adequate manier om populatieparameters te schatten. Het bestaan van heteroscedasticiteit kan eveneens worden afgeleid uit de scatterplot, waarbij de gestandaardiseerde residuele termen worden uitgezet als functie van de voorspelde waarden. We krijgen een patroon te zien van uitwaaierende residuele termen, m.a.w. er is geen random spreiding van de errortermen rond de verwachte waarde van Y. Hoe groter de verwachte waarde van Y, hoe groter het verschil tussen de geobserveerde waarde voor Y en de voorspelde waarde voor Y. Een manier om hiermee om te gaan is het gebruik van WLS-regressie, waarbij men een verschillend gewicht geeft aan de verschillende observaties, doch dit valt buiten het bestek van deze inleidende cursus.



### Additiviteit

We vervolgen met een controle op **additiviteit**. We gaan hierbij als volgt te werk: een model wordt opgesteld dat bestaat uit alle mogelijke interactie-effecten en wordt vergeleken met een statistisch model met enkel de hoofdeffecten. Indien een regressiemodel met interactie een significante verbetering teweeg brengt in de verklaarde variantie, is dit een indicatie voor het feit dat niet elke variabele op zich een effect heeft op de criminaliteitsgraad, maar dat de verschillende variabelen elkaar versterken in hun effect op de graad. We spreken dan van conditionaliteit.

---

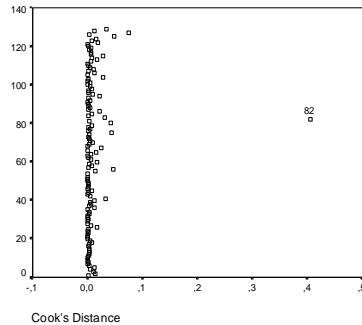
<sup>17</sup> Mc Clendon, J. (1994). *Multiple regression and Causal analysis*. New York: Peacock Publishers.

## Lineariteit

Lineariteit betekent dat het verband tussen X en Y lineair is. Wanneer we geen speciaal curvilineair patroon ontwaren, behouden we de lineaire regressie. In het geval er curvilineariteit bestaat, moeten we curvilineaire regressietechnieken gebruiken. Deze vallen buiten het bestek van deze cursus.

## Uitbijters of outliers

Tot slot is het belangrijk na te gaan of er in een regressiemodel **uitbijters** te detecteren vallen die meer dan twee standaardafwijkingen boven de gestandaardiseerde verwachte waarde liggen. Uitbijters zijn waarden die onmiddellijk in het oog springen omwille van hun geïsoleerde positie in de grafische weergave van de analyse. Het gaat om observaties met een wel erg groot verschil tussen de geobserveerde waarde en de voorspelde waarde op basis van een onafhankelijke variabele. De aanwezigheid van uitbijters hoeft op zich geen probleem te zijn. Het is perfect denkbaar dat bijvoorbeeld een bepaalde gemeente in een regressieanalyse een eenzame positie heeft, doch dit wil nog niet zeggen dat deze waarde de regressielijn zwaar beïnvloedt. Het probleem doet zich pas voor wanneer deze waarde de regressielijn extreem beïnvloedt. Dit kunnen we nagaan door **Cook's distance** op te vragen in een software programma zoals SPSS. Traditioneel wordt aangenomen dat waarden groter dan 1 een eerder problematisch karakter kennen.



## 9. De limieten van meervoudige regressie

Er wordt te vaak van uitgegaan dat de meervoudige regressie het probleem van de identificatie van de unieke causale impact van oorzaken overwint en ons in staat stelt hun relatieve belang ten aanzien van diverse uitkomstvariabelen af te wegen. Heel vaak moeten we vaststellen dat ook voorzichtige auteurs de volgende ideeën door elkaar halen:

1. A (een ‘variabele’, vb. IQ) correleert met B (een andere variabele, vb. inkomen)
2. A ‘voorspelt’ B
3. A ‘verklaart de variantie’ in B
4. A ‘verklaart’ B
5. A ‘veroorzaakt’ B

We kunnen punten 4 en 5 samen behandelen. We kunnen ‘A verklaart B’ zeggen, alleen als we kunnen zeggen dat ‘A veroorzaakt B’. Maar, **ten eerste**, correlaties staven geen oorzaken. Oorzaken zijn ‘mechanismen’ die uitkomsten produceren. Zo kunnen we correlaties hebben waarbij er geen mechanisme denkbaar is, vb. tussen de prijs van eieren op een markt in Beijing en de prijs van Microsoft op de New Yorkse aandelenbeurs. **Ten tweede**, er bestaan heel veel oorzaken van een uitkomst. Zo is er, wanneer we vuur willen maken, zowel brandbaar materiaal nodig als een bron van warmte en zuurstof. Als er één van de voornoemde variabelen ontbreekt, kan er geen vuur zijn. Welke is nu meer belangrijk? Wel, we zullen enkel vuur krijgen bij een juiste combinatie. (Om een vinylen stof te laten ontbranden heb je meer warmte nodig dan voor het ontbranden van katoen). Als we een bron van warmte nemen als ‘de oorzaak’, is dat omdat we ervan uitgaan dat er zuurstof en brandbaar materiaal aanwezig zijn. We vergeten eigenlijk de zuurstof en zeggen dat een vonk het vuur heeft ‘veroorzaakt’. Weber noemde dit ‘*adequate causation*’, het verschil in de bestaande staat die het effect met zich meebracht. Dit is pragmatisch en niet verrassend, maar het blijft wel een feit dat alle factoren belangrijk zijn: je zal geen vuur hebben als er een factor ontbreekt. Neem nu de mogelijkheid van studenten uit de eerste BAC criminologie om hoog te scoren op statistiek. Wat is ‘de oorzaak’? Welke van deze ‘factoren’ (oorzaken) zijn het belangrijkste? De ene student kan ‘slim’ zijn, maar de ander kan ook enorm gemotiveerd zijn, goed les gekregen hebben, een andere kan zich goed gevoeld hebben op de dag van de test. Met multivariate regressiemodellen tonen we aan welke effecten het sterkst samenhangen in

steekproeven of in populaties (wanneer we populatiegegevens hebben). Kortom, denken in termen van lineaire en additieve invloeden is niet steeds correct. Er moet gezocht worden naar relevante causale elementen die samen een voldoende oorzaak vormen voor de verklaring van een effect via het in gang zetten van een causaal mechanisme.

## 10. Testvragen

Hieronder vind je enkele uitspraken over de multivariate statistiek. Deze vragen kan je gebruiken om je parate kennis te toetsen over de multivariate analyse en pad-analyse. **In gewijzigde vorm kunnen dergelijke theorievragen ook op het examen voorkomen.** Deze vragen zijn afkomstig uit vroegere examens.

- 1. De meervoudige regressieanalyse met X1 en X2 als onafhankelijke variabelen geeft hetzelfde resultaat als twee afzonderlijke bivariate regressieanalyses. Klopt dat?**
  - Dit klopt enkel in de situatie waarbij  $r(x_1, x_2) = 0$ .
  - Dit klopt nooit
- 2. R (hoofdletter) staat in de output van een multiplele regressie voor de correlatie tussen Y en de verwachte waarde voor Y op basis van de onafhankelijke variabelen**
  - Deze uitspraak is juist
  - Deze uitspraak is verkeerd
- 3. Een meervoudige multivariate regressieanalyse is een**
  - Regressieanalyse met meerdere onafhankelijke variabelen en meerdere afhankelijke variabelen
  - Regressieanalyse met meerdere onafhankelijke variabelen en slechts een afhankelijke variabele
  - Regressieanalyse met meerdere afhankelijke variabele en een onafhankelijke variabele

**4. Heteroscedasticiteit wil zeggen dat**

- De waarde van de residuele termen toenemen naarmate X1 toeneemt
- De waarde van de residuele termen gelijk blijft, naarmate X1 toeneemt

**5. Additiviteit betekent dat X1 en x2 onafhankelijke effecten hebben, ttz ze dragen elk bij tot de verklaring van Y**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**6. Lineariteit wil zeggen dat Y als lineaire functie van X kan worden uitgedrukt**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**7. Curvilineariteit wil zeggen dat het effect van X1 op Y toeneemt of afneemt naargelang X1 toeneemt.**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**8. Een uitbijter kan de regressielijn ernstig beïnvloeden**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**9. Interactie betekent dat het effect van X1 op Y conditioneel is op X2**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**10. Een mediatorvariabele is een variabele die het effect van een exogene variabele op een afhankelijke variabele medieert (dwz dat deze variabele het effect van de exogene variabele wegverklaart en dat het effect van de exogene variabele op Y via de mediator verloopt)**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**11. Een totaal effect is in een pad-analyse gelijk aan de rechtstreekse effecten min de onrechtstreekse effecten**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

## **11. Leerdoelen**

De doelstelling van dit hoofdstuk ligt in het begrijpen en kunnen toepassen van de meervoudige regressieanalyse. Studenten dienen de meervoudige regressieanalyse zelf te kunnen toepassen. Deze inleiding is beperkt in die zin dat we enkel voorbeelden hebben gegeven van meervoudige regressieanalyses met twee onafhankelijke variabelen. Dit betekent dat studenten de parameters van de meervoudige regressieanalyse in de situatie met twee onafhankelijke variabelen zelf moeten kunnen uitrekenen. Het is cruciaal dat de studenten snappen waarom we niet zomaar twee afzonderlijke bivariate analyses bij elkaar kunnen optellen. De meervoudige regressieanalyse is echter een techniek die heel wat eisen stelt aan de data. Deze eisen zijn velerlei en werden slechts partieel besproken. Wat belangrijk is, is dat de lineaire OLS-regressieanalyse niet steeds bruikbaar is. In het geval van ernstige schendingen van de veronderstellingen (heteroscedasticiteit, normaliteit, lineariteit,...) moeten andere methoden gekozen worden. Deze andere methoden zijn ook regressieanalyses, maar deze zijn niet gebaseerd op het OLS-principe. Echter, de inhoudelijke interpretatie van de regressiecoëfficiënten is analoog. Deze andere varianten komen in dit handboek niet aan bod, maar wel in vervolg cursussen. Voorbeelden zijn de logistische regressie, de Poisson-regressie en de negatief-binomiaal regressie. Deze technieken stellen minder strenge eisen aan de structuur van de data. Zo heeft de scheefheid van de verdeling minder een invloed op de berekening van de parameters en de significantie van deze parameters in de laatst genoemde technieken. Tot slot zijn we heel concreet ingegaan op de statistische interactie door een

voorbeeld te geven van een analyse waarbij het verband tussen twee metrische kenmerken conditioneel is op een derde kenmerk.

## Hoofdstuk 12

### Complexere relaties tussen variabelen

#### 1. Inleidende begrippen

*Onafhankelijke, intermediaire en afhankelijke variabelen*

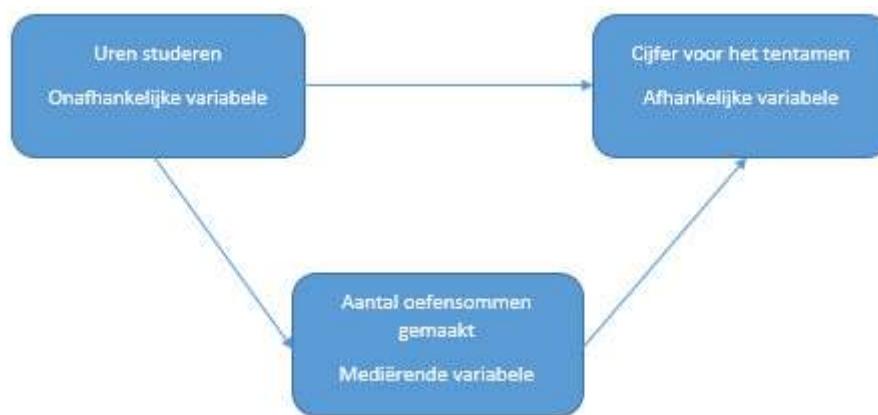
Bij een oorzaak-gevolgrelatie is er altijd sprake van een relatie tussen minstens twee variabelen en een mechanisme dat oorzaak en gevolg met elkaar verbindt. Onafhankelijke variabelen worden in de causale variabelentaal exogene variabelen genoemd. Zij hebben effecten op de endogene variabelen. Endogene variabelen zijn afhankelijke variabelen. In de pad-analyse zijn er meerdere afhankelijke variabelen en is het zo dat een afhankelijke variabele in een deel van het statistisch model ook een effect heeft op andere variabelen in het statistisch model.

Wanneer we de relatie tussen drie of meerdere variabelen bestuderen, spreken we van netwerken van relaties tussen variabelen. De concepten en hun onderlinge relaties kunnen worden samengevat in een **conceptueel model**. Dit wordt soms ook een **conceptueel diagram** genoemd. Wanneer de relaties tussen de variabelen op basis van empirisch onderzoek tot stand komen, spreken we van een **padmodel of een paddiagram**. Het begrip 'structureel model' wordt eveneens gebruikt om deze modellen aan te duiden. Zowel in het conceptuele model als in eenvoudige padmodellen zijn er **3 basisposities voor de variabelen**. **Helemaal rechts** staan de variabelen die verklaard worden door andere variabelen. Dit zijn de **afhankelijke variabelen**. **Helemaal links staan de onafhankelijke variabelen**. Dit zijn de **variabelen vanwaar het effect uitgaat**. Tussen deze twee types variabelen in staan **de intermediaire variabelen**. Dit zijn variabelen die verklaard worden door één of verscheidene onafhankelijke variabelen en die tevens verklarend zijn voor de afhankelijke variabelen.

#### 2. Mediatorvariabele

Een mediërende variabele of mediatorvariabele is een variabele die tussen een oorzaak-gevolgrelatie staat en hiermee het effect tussen de oorzaak-gevolgrelatie beter verklaart en de relatie sterker maakt. De mediatorvariabele noemt men vaak een mechanisme waarlangs een externe oorzaak (of beter gezegd: de eigenschap van een gebeurtenis) het gevolg teweegbrengt.

In deze paragraaf gaan we door op het voorbeeld van de student die veel studeert (onafhankelijke variabele) om zo een hoger cijfer te halen voor zijn tentamen (afhankelijke variabele). We voegen nu een mediërende variabele aan het voorbeeld toe. De mediërende variabele staat tussen de onafhankelijke en afhankelijke variabele in (de oorzaak-gevolgrelatie) en stelt je in staat om de oorzaak-gevolgrelatie beter te verklaren. Een mediërende variabele kan erg lastig zijn om te interpreteren en om conclusies aan te verbinden. Daarom onderbouw je een mediërende variabele altijd met een statistische analyse. Zie hieronder een voorbeeld van een mediërende variabele.



### Voorbeeld conceptueel model met mediërende variabele

In dit voorbeeld staat de relatie tussen de onafhankelijke variabele ‘Uren studeren’ en de afhankelijke variabele ‘Cijfer voor het tentamen’ centraal. De oorzaak-gevolgrelatie is dat hoe meer uren de student studeert, hoe hoger het cijfer is voor het examen. Nu hebben we de mediërende variabele ‘Aantal oefensommen gemaakt’ toegevoegd. Zoals je in het voorbeeld kunt zien, staat de mediërende variabele tussen de onafhankelijke en afhankelijke variabele. Hoe meer uren een student studeert, hoe meer oefensommen de student maakt en hoe meer oefensommen de student maakt, hoe hoger het cijfer voor het tentamen. Door de mediërende variabele ‘Aantal oefensommen gemaakt’ toe te voegen versterken we de oorzaak-gevolgrelatie tussen de onafhankelijke en afhankelijke variabele.

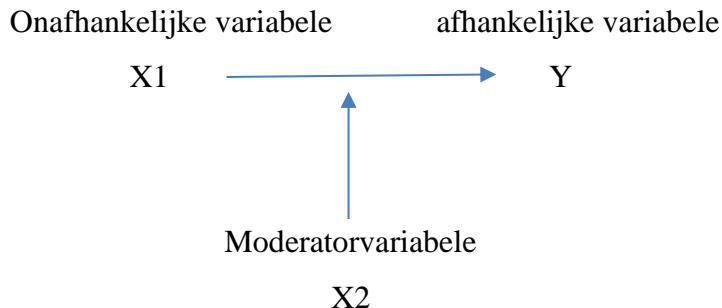
### 3. Moderatorvariabele of het interactie-effect

#### *Inleiding*

In kwantitatief onderzoek verwijst de term ‘interactie’ naar de situatie waarbij het effect van een onafhankelijke variabele X<sub>1</sub> op een afhankelijke variabele Y beïnvloed wordt door een derde onafhankelijke variabele X<sub>2</sub>. Deze laatste variabele wordt de **moderatorvariabele**

genoemd omdat deze variabele de wijze verandert waarop X1 een impact heeft op de afhankelijke variabele Y. Figuur 1 toont hoe een interactie-effect visueel wordt voorgesteld.

*Figuur 1: schematische voorstelling van een interactie-effect*



In criminologisch empirisch onderzoek is het belangrijk dat de onderzoeker op voorhand een bepaalde theoretische verwachting heeft. We bedoelen hiermee dat de onderzoeker, op basis van een theorie of voorgaand empirisch onderzoek, een bepaalde veronderstelling heeft van welke X-variabele in eerste instantie een effect heeft op een Y-variabele en van welke X-variabele verwacht wordt dat die de relatie wijzigt, dus welke X-variabele potentieel een moderatorvariabele is. Bijvoorbeeld: volgens sociale leertheorieën hebben delinquente vrienden een positief effect op het plegen van delicten bij adolescenten. Schematisch kunnen we deze relatie als volgt noteren:

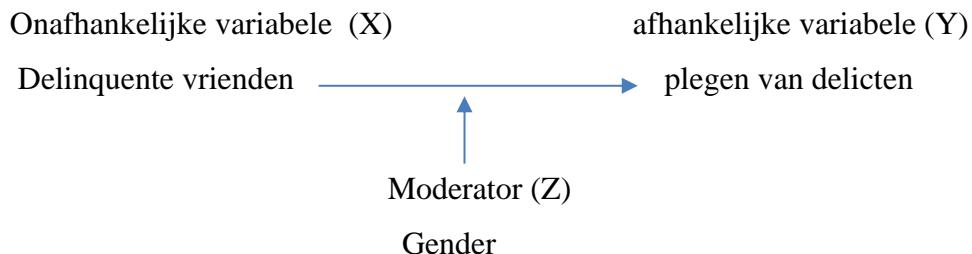
*Figuur 2: schematische voorstelling van het hoofdeffect van delinquente vrienden op het plegen van delicten*



We verwachten op basis van de sociale leertheorie in de eerste plaats een **hoofdeffect** van delinquente vrienden op het plegen van delicten;

Op basis van deze theoretische veronderstelling (die trouwens in heel veel criminologisch empirisch onderzoek wordt gerepliceerd) zouden we ons kunnen afvragen of de impact van delinquente vrienden op het plegen van delicten voor *elke* adolescent hetzelfde is. Wordt elke adolescent *op dezelfde manier* beïnvloed door delinquente vrienden? Zou het bijvoorbeeld niet kunnen dat er een verschil is tussen jongens en meisjes? Deze vraag stellen betekent dat we ons afvragen of de variabele 'gender' (meisje of jongen-zijn) een impact heeft op de

relatie tussen delinquenten vrienden en het plegen van delicten. Als de effecten van delinquenten vrienden op het plegen van delicten inderdaad voor jongens en meisjes anders zijn, dan spreekt men van *interactie*. Schematisch kunnen we dit voorstellen als volgt:



De verwachting is dat de relatie delinquenten vrienden en het plegen van delicten anders is voor jongens dan voor meisjes.<sup>1</sup> Als we verwachten dat het effect van delinquenten vrienden op het plegen van delicten verschillend is voor meisjes en voor jongens dan beschouwen we ‘gender’ als een **moderatorvariabele**. Dus nogmaals: in het geval van interactie onderzoeken we of het effect van X op Y verandert als Z een andere waarde heeft (dat wil zeggen: andere groep, categorie of conditie).

### ***Enkele voorbeelden van interactie-effecten tussen variabelen op laag meetniveau***

Hierna geven we twee concrete voorbeelden van statistische interacties aan de hand van kruistabellen. De voorbeelden zijn afkomstig uit voorgaand criminologisch onderzoek.<sup>2</sup>

In het eerste voorbeeld kijken we naar de relatie tussen immigratie-achtergrond (nominale variabele met 2 categorieën: Belg of immigrant) en lidmaatschap van een problematische jeugdgroep (nominale variabele met 2 categorieën : wel of geen lidmaatschap), dus twee variabelen van het nominale meetniveau.

---

<sup>1</sup> In dit geval formuleren we onze verwachting nog zeer algemeen zonder een uitspraak te doen over de sterkte van het hoofdeffect. Hou steeds voor ogen dat verwachtingen of veronderstellingen die onderzoekers uitspreken altijd gebaseerd zijn op theorie of voorgaand onderzoek. De keuze voor een moderatorvariabele is immers niet altijd eenduidig. Wat voor de ene onderzoeker een moderatorvariabele is, is dat soms niet voor een andere onderzoeker. Vandaar dat het belangrijk is om de keuze goed theoretisch te onderbouwen (of op basis van eerder empirisch onderzoek).

<sup>2</sup> Zie voor meer informatie: PAUWELS, L., & HARDYNS, W. (2016). *Problematic youth group involvement as situated choice: testing an integrated conditions-controls-exposure model*. Eleven International Publishing.

*Tabel 1: Kruistabel lidmaatschap problematische jeugdgroep naar immigratie-achtergrond*  
**Lidmaatschap problematische jeugdgroep naar immigratie-achtergrond**

PROBLEMATISCHE JEUGDGROEP	GEEN LID	Immigratie-achtergrond			Totaal
		Belg	Immigrant		
		Absolute aantal	1041	1235	2276
	WEL LID	Kolompercentage	95.1%	89.4%	91.9%
		Absolute aantal	54	146	200
Total		Kolompercentage	4.9%	10.6%	8.1%
		Absolute aantal	1095	1381	2476
		Kolompercentage	100.0%	100.0%	100.0%

$\chi^2 = 26.170$ ; df = 1;  $p < .001$

Tabel 1 is een kruistabel die de samenhang weergeeft tussen 2 nominale variabelen. In de kolommen staat de variabele ‘immigratie-achtergrond’ die bestaat uit 2 categorieën: Belg en immigrant. In de rijen staat de variabele ‘problematische jeugdgroep’ die ook bestaat uit 2 categorieën: geen lid en wel lid. Herinner je uit hoofdstuk 6 (*bivariate associatiematen voor nominale en ordinale variabelen*) dat we bij het maken van een bivariate kruistabel statistisch een onderscheid maken tussen een onafhankelijke en een afhankelijke variabele en dat de richting van de associatie bepaald wordt door onze theoretische verwachtingen. Herinner je ook dat we in deze cursus de afspraak maakten om de variabele die we beschouwen als onafhankelijke variabele in de kolommen te plaatsen en de afhankelijke variabele in de rijen. Elke cel geeft ons informatie over de twee variabelen. Kijken we naar de conditionele verdelingen (hoe is één van de twee variabelen verdeeld binnen één categorie van de andere variabele) dan zien we voor de categorie Belg dat 95,1% van de respondenten rapporteerden GEEN LID te zijn van een problematische jeugdgroep en 4,9% rapporteerden WEL LID te zijn. Voor de categorie Immigrant rapporteerden 89,4% van de respondenten GEEN LID te zijn van een problematische jeugdgroep en 10,6% rapporteerden WEL LID te zijn. We besluiten dat de frequentieverdeling voor de beide categorieën van de variabelen niet dezelfde is: er is een verschil tussen Belgen en Immigranten in termen van hun betrokkenheid bij een problematische jeugdgroep. Het verband tussen beide variabelen is bovendien significant ( $\chi^2=26.170$ ;  $p < .001$ ).

Om na te gaan of de samenhang tussen immigratie-achtergrond en lidmaatschap problematische jeugdgroep verschillend is voor meisjes en voor jongens kijken we naar tabel 2. Deze kruistabel toont een aantal verschillen. Uit tabel 2 blijkt dat van de Belgische meisjes slechts 3.6% aangaf WEL LID te zijn van een problematische jeugdgroep. Bij de Belgische jongens ligt de proportie duidelijk hoger, respectievelijk 6.4%. Kijken we naar de categorie Immigrant dan merken we dat binnen deze groep 6.9% van de meisjes aangeeft WEL LID te zijn van een problematische jeugdgroep, terwijl deze proportie bij de jongens ook duidelijk hoger ligt, namelijk 14.2% van de Immigrant-jongens geeft aan WEL LID te zijn van een problematische jeugdgroep.

We merken dus duidelijk dat de relatie tussen immigratie-achtergrond en lidmaatschap problematische jeugdgroep verschilt tussen de geslachten. Er zijn meer jongens, zowel in de categorie Belg als Immigrant, die WEL LID zijn van een problematische jeugdgroep, in vergelijking met Belgische meisjes en Immigrant-meisjes.

*Tabel 2: kruistabel lidmaatschap problematische jeugdgroep naar immigratie-achtergrond en geslacht*

**Lidmaatschap problematische jeugdgroep naar immigratie-achtergrond en geslacht**

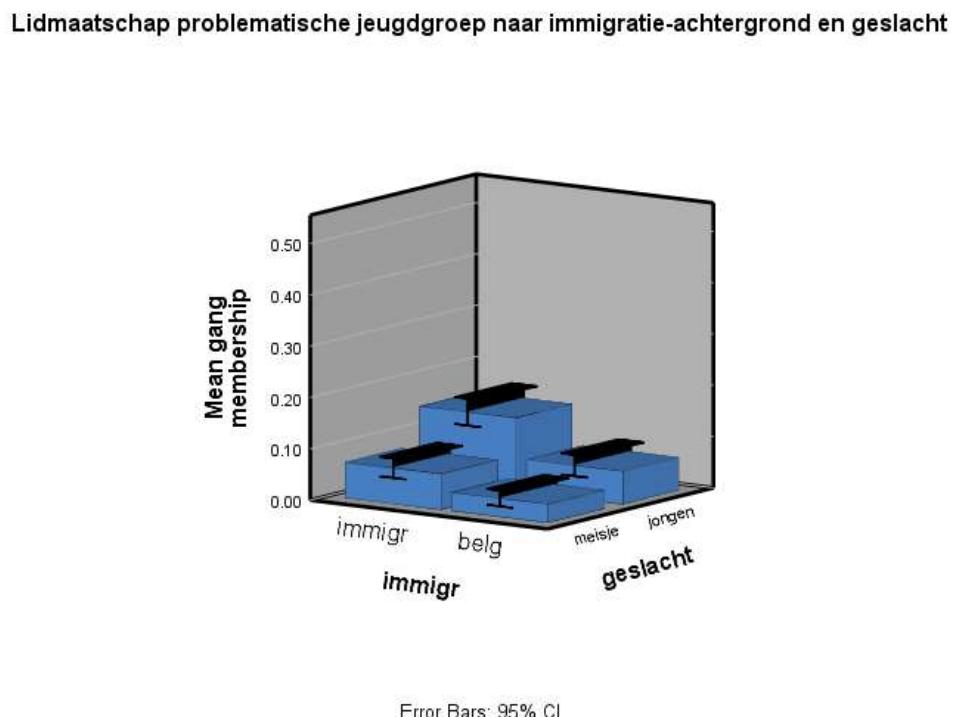
GESLACHT			Immigratie-achtergrond		
			Belg	immigrant	Totaal
MEISJE	PROBLEMATISCHE JEUGDGROEP	GEEN LID	Absolute aantal	541	644
			kolompercentage	96.4%	93.1%
	WEL LID	Absolute aantal	20	48	68
		kolompercentage	3.6%	6.9%	5.4%
JONGEN	PROBLEMATISCHE JEUGDGROEP	GEEN LID	Absolute aantal	561	692
			kolompercentage	100.0%	100.0%
	WEL LID	Absolute aantal	34	98	132
		kolompercentage	6.4%	14.2%	10.8%
	Totaal	Absolute aantal	534	688	1222
		kolompercentage	100.0%	100.0%	100.0%

Meisjes: Chi<sup>2</sup>= 6,861; df=1; p<.01

Jongens: Chi<sup>2</sup>=19,361; df=1; p<.001

In figuur 2 presenteren we de resultaten aan de hand van een drie-dimensionele bar chart waarbij de proportie respondenten die aangeeft WEL LID te zijn van een problematische jeugdgroep duidelijk kan worden afgelezen voor de twee categorieën immigratie-achtergrond en voor zowel jongens als voor meisjes. De figuur spreekt voor zich: de proportie respondenten die aangeeft WEL LID te zijn van een problematische jeugdgroep is het laagst bij de Belgische meisjes en het hoogst bij de immigrant-jongens.

*Figuur 2: 3D-voorstelling van interactie tussen immigratie-achtergrond en geslacht*



We presenteren een tweede voorbeeld in tabel 3.

In eerste instantie onderzoeken we de relatie tussen de variabelen ‘veelpleger’, die we definiëren als respondenten die minstens 4 gepleegde feiten rapporteren, en gezinsstructuur (bestaande uit twee categorieën: éénouder en twee-oudergezin).

Uit de resultaten blijkt dat 87.6% van de respondenten die opgroeien in een éénoudergezin aangeven ‘GEEN veelpleger’ te zijn tegenover 12.4% die aangeeft ‘WEL veelpleger’ te zijn. Binnen de categorie ‘twee-oudergezin’ zijn de proporties respectievelijk 82.1% (GEEN veelpleger) en 17.9% (WEL veelpleger). We zien dus dat de frequentieverdeling voor de beide categorieën van de variabelen niet dezelfde is: er is een verschil tussen respondenten die opgroeien binnen een éénouder- of twee-oudergezinsstructuur met betrekking tot GEEN/WEL veelpleger. Het verband tussen beide variabelen is significant ( $\chi^2 = 8.091$ ;  $p < .01$ ).

Tabel 3: Kruistabel veelpleger en gezinsstructuur

**Veelpleger naar gezinsstructuur**

VEELPLEGER (minstens 4 feiten)			gezinsstructuur			Totaal	
			Woont in éénouder gezin	Woont in twee-ouder gezin			
GEEN		Absolute aantal	1818	299	2117		
		kolompercentage	87.6%	82.1%	86.8%		
WEL		Absolute aantal	257	65	322		
		kolompercentage	12.4%	17.9%	13.2%		
Total		Absolute aantal	2075	364	2439		
		kolompercentage	100.0%	100.0%	100.0%		

Chi<sup>2</sup> = 8.091; df=1; p<.01

In tweede instantie vragen we ons af of het verband tussen ‘gezinsstructuur’ en ‘veelpleger’ verschilt naargelang de mate van familiale controle. In tabel 4 presenteren we de frequentieverdelingen van de categorieën gezinsstructuur voor GEEN/WEL veelpleger en LAGE/HOGE familiale controle.

Wat kunnen we aflezen uit de proportionele verdeling ?

Tabel 4: Kruistabel veelpleger naar gezinsstructuur (éénouder – twee-ouders) en familiale controle (laag – hoog)

**Veelpleger naar gezinsstructuur en familiale controle**

			Gezinsstructuur		
			Woont in éénouder gezin	Woont in twee-ouder gezin	Totaal
FAMILIALE CONTROLE	LAGE FAMILIALE CONTROLE	VEELPLEGER (minstens 4 feiten)	GEEN	Absolute aantal kolompercentage	997 82.0%
			VEELPLEGER	Absolute aantal kolompercentage	219 18.0%
					57 26.4% 276 19.3%
	Total			Absolute aantal kolompercentage	1216 100.0%
					216 100.0% 1432 100.0%
HOGE FAMILIALE CONTROLE	HOGE FAMILIALE CONTROLE	VEELPLEGER (minstens 4 feiten)	GEEN	Absolute aantal kolompercentage	797 95.9%
			VEELPLEGER	Absolute aantal kolompercentage	34 4.1%
					7 4.8% 41 4.2%
	Total			Absolute aantal kolompercentage	831 100.0%
					145 100.0% 976 100.0%

Lage familiale controle: Chi<sup>2</sup>= 8,277; df=1; p<.01

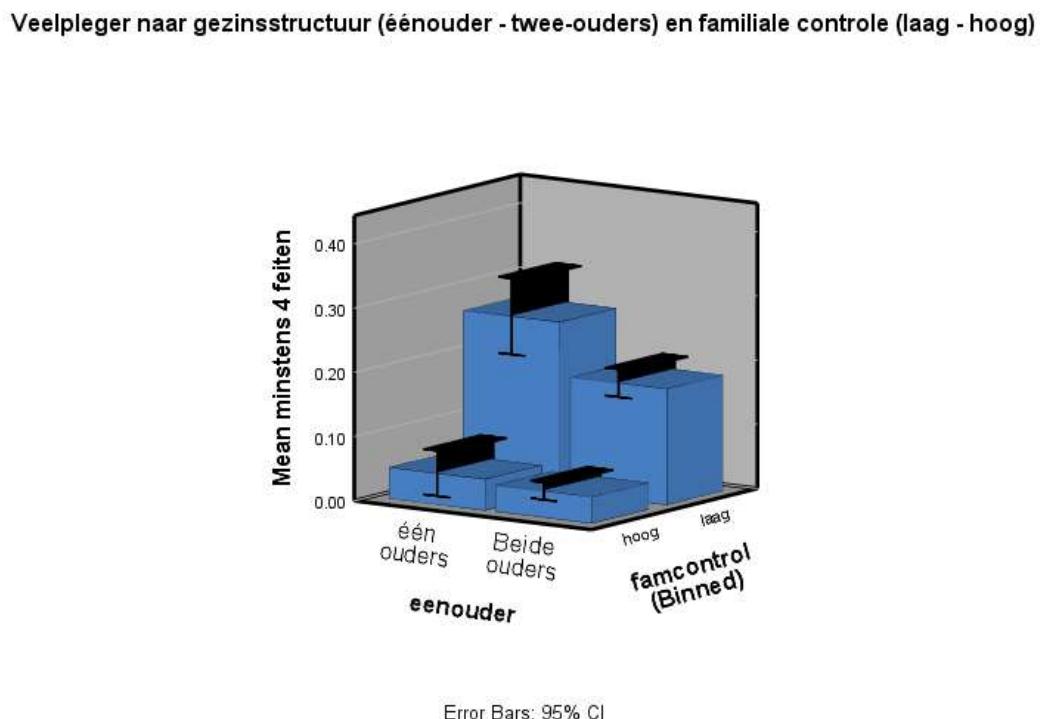
Hoge familiale controle: Chi<sup>2</sup>= 0,166; df = &; p=.683

Uit tabel 4 lezen we af dat van de respondenten die opgroeien binnen een éénoudergezin met een lage familiale controle 18.0% veelpleger blijkt te zijn. Deze proportie ligt veel lager voor respondenten uit een éénoudergezin met hoge familiale controle, namelijk 4.1%. Kijken we naar de categorie twee-oudergezin dan merken we dat binnen deze groep 26.4% veelpleger is wanneer er sprake is van een lage familiale controle terwijl slechts 4.8% van de respondenten veelpleger is wanneer er sprake is van hoge familiale controle. We merken duidelijk dat er verschillen zijn in de samenhang tussen gezinsstructuur en veelpleger naargelang de mate van familiale controle.

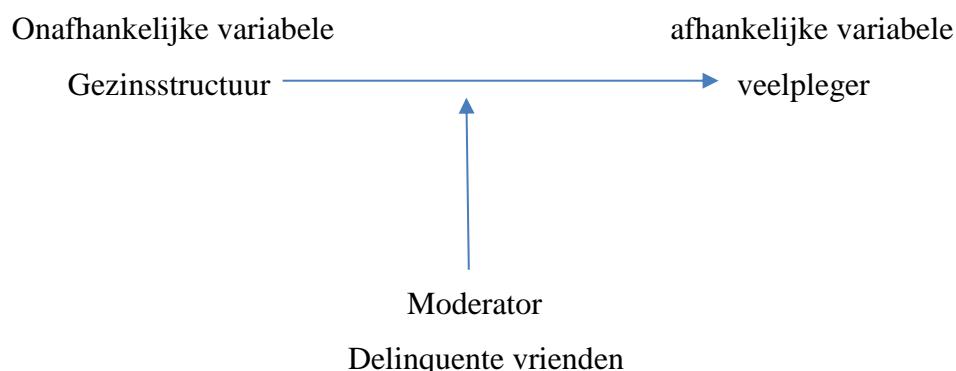
In figuur 3 presenteren we terug de resultaten aan de hand van een drie-dimensionele bar chart waarbij de proportie respondenten die aangeeft veelpleger te zijn duidelijk kan worden afgelezen voor de twee categorieën gezinsstructuur en mate van familiale controle(hog/laag).

De visuele voorstelling is zeer duidelijk: de proportie respondenten die aangeeft veelpleger te zijn is het laagst binnen de één- en tweeouder gezinsstructuur met hoge familiale controle en overduidelijk het hoogst binnen de categorie éénoudergezin met een lage familiale controle.

*Figuur 3: 3D-voorstelling van interactie tussen gezinsstructuur en familiale controle*



Tot slot nog een laatste voorbeeld met een moderatorvariabele van het ordinale meetniveau: delinquente vrienden (laag – medium – hoog). We onderzoeken de relatie tussen gezinsstructuur, delinquente vrienden en veelpleger. Laat ons de veronderstelde samenhang eerst conceptueel voorstellen. We veronderstellen een hoofdeffect van gezinsstructuur op de afhankelijke variabele veelpleger. Daarnaast verwachten we dat deze samenhang verschillend zal zijn naar de mate waarin respondenten delinquente vrienden hebben. We veronderstellen dat delinquente vrienden een moderator is in de relatie tussen gezinsstructuur en veelpleger.



Tabel 5 presenteert de frequentieverdelingen van de twee categorieën gezinsstructuur (éénouder en twee-ouders) naar GEEN/WEL veelpleger en drie categorieën delinquenten vrienden (laag – medium – hoog). Helemaal onderaan de kruistabel lezen we de totale frequenties af voor wat betreft de samenhang tussen gezinsstructuur en veelpleger (totaal).

12.1% van de respondenten die opgroeien in een tweeoudergezin zijn veelpleger tegenover 17.8% van de respondenten die opgroeien in een éénoudergezin. Het verband tussen de variabelen ‘veelpleger’ en ‘gezinsstructuur’ is significant ( $\text{Chi}^2 = 8.743$ ;  $p < .01$ ). Vervolgens gaan we na of de samenhang tussen gezinsstructuur en veelpleger verschillend is naargelang het hebben van delinquenten vrienden. ‘Delinquenten vrienden’ is een variabele van het ordinale meetniveau en bestaat uit drie categorieën: laag – medium – hoog. Voor wat betreft de categorie: laag niveau van delinquenten vrienden zien we geen verschil tussen respondenten uit een éénouder of tweeoudergezin met betrekking tot veelpleger zijn (resp. 1.2% en 1.3%). Voor de categorie: medium niveau van delinquenten vrienden zien we dat 7.0% van de respondenten uit een tweeoudergezin veelpleger zijn tegenover 10.4% van de respondenten uit een éénoudergezin. Het verschil bedraagt hier 3.4 procentpunten. In de categorie: hoog niveau van delinquenten vrienden is het verschil tussen veelpleger in een éénouder- en een tweeoudergezin 4.1 procentpunten (resp. 42.9% en 38.8%). Besluit: we merken dat de samenhang tussen gezinsstructuur en veelpleger verschillend is naar de mate waarin respondenten delinquenten vrienden hebben.

Tabel 5 : Kruistabel veelpleger naar delinquenten vrienden en gezinsstructuur

**Veelpleger naar delinquenten vrienden (laag – medium – hoog) en gezinsstructuur (éénouder – tweeouders)**

DELINQUENTE VRIENDEN	VEELPLEGER	GEEN VEELPLEGER	Gezinsstructuur			Totaal
			Woont in Twee- oudergezin	Woont in één oudergezin		
Delinquenten vrienden : LAAG	VEELPLEGER (minstens 4 feiten)	GEEN VEELPLEGER	Absolute aantal	468	83	551
			kolompercentage	98.7%	98.8%	98.7%
		VEELPLEGER	Absolute aantal	6	1	7
			kolompercentage	1.3%	1.2%	1.3%
	Total		Absolute aantal	474	84	558
			kolompercentage	100.0%	100.0%	100.0%
Delinquenten vrienden: MEDIUM	VEELPLEGER (minstens 4 feiten)	GEEN VEELPLEGER	Absolute aantal	1057	147	1204
			kolompercentage	93.0%	89.6%	92.6%
		VEELPLEGER	Absolute aantal	79	17	96
			kolompercentage	7.0%	10.4%	7.4%
	Total		Absolute aantal	1136	164	1300
			kolompercentage	100.0%	100.0%	100.0%
Delinquenten vrienden: HOOG	VEELPLEGER (minstens 4 feiten)	GEEN VEELPLEGER	Absolute aantal	252	60	312
			kolompercentage	61.2%	57.1%	60.3%
		VEELPLEGER	Absolute aantal	160	45	205
			Kolompercentage	38.8%	42.9%	39.7%
	Total		Absolute aantal	412	105	517
			kolompercentage	100.0%	100.0%	100.0%

Delinquenten vrienden laag: Chi<sup>2</sup>= 0,003; df=1; p= .954

Delinquenten vrienden medium: Chi<sup>2</sup>= 2,439; df=1; p= .118

Delinquenten vrienden hoog: Chi<sup>2</sup>= 0,566; df=1; p=.452

### ***Interactie in een meervoudige lineaire regressie***

Het gebruik van kruistabellen met verschillende niveaus om interactie-effecten op te sporen heeft beperkingen wanneer we het effect van meerdere onafhankelijke variabelen op één afhankelijke variabele onderzoeken, zeker als één of meerdere onafhankelijke variabelen van het metrisch meetniveau zijn.

Een regressiemodel laat toe het effect van verschillende onafhankelijke variabelen op één afhankelijke variabele gelijktijdig te onderzoeken. Als er twee onafhankelijke variabelen zijn dan is de regressievergelijking als volgt:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Herhaling van de algemene interpretatie van de regressievergelijking:

$\hat{Y}$  is de voorspelde waarde van de afhankelijke variabele,  $\alpha$  is het intercept of de voorspelde waarde voor de afhankelijke variabele als alle onafhankelijke variabelen gelijk zijn aan 0. De richtingscoëfficiënten of regressiecoëfficiënten ( $\beta_1$  en  $\beta_2$ ) geven het effect aan van de onafhankelijke variabelen  $X_1$  en  $X_2$  op de afhankelijke variabele. In een meervoudige regressievergelijking (met meerdere onafhankelijke variabelen) is dat het effect van  $X$  op  $Y$  gecontroleerd voor de effecten van alle andere onafhankelijke variabelen in het model. Letterlijk is  $\beta_1$  dan de voorspelde wijziging in  $Y$  bij één eenheid wijziging in  $X_1$  onder controle van de effecten van  $X_2$  op  $Y$ .

Om een interactie-effect in een regressiemodel te schatten wordt in het model een productterm van de interagerende variabelen opgenomen. Als er twee onafhankelijke variabelen zijn, dan ziet de vergelijking eruit als volgt:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

In deze vergelijking worden  $\beta_1$  en  $\beta_2$  wel eens hoofdeffecten (*main effects*) genoemd en  $\beta_3$  het interactie effect.

De interpretatie van het intercept is in deze vergelijking dezelfde: de voorspelde waarde voor de afhankelijke variabele als alle onafhankelijke variabelen gelijk zijn aan 0.

$\beta_1$  is het effect van  $X_1$  op  $Y$  als  $X_2$  gelijk is aan 0 en  $\beta_2$  is het effect van  $X_2$  op  $Y$  als  $X_1$  gelijk is aan 0. Het is dus duidelijk dat het effect van  $X_1$  op  $Y$  afhankelijk is van  $X_2$  en omgekeerd. Tot slot:  $\beta_3$  geeft het effect weer van de interactieterm.

Laat ons een voorbeeld geven van een meervoudig regressiemodel met als afhankelijke variabele zelf gerapporteerde criminaliteit en als onafhankelijke variabelen geslacht, immigratieachtergrond en cohorte (als controlevariabelen), criminale geneigdheid en lifestyle risk en de interactieterm tussen deze laatste als onafhankelijke variabelen.

*Tabel 6: regressiemodel met zelf gerapporteerde criminaliteit als afhankelijke variabele en geslacht, immigratie achtergrond, cohorte, criminale geneigdheid, lifestyle risk en de interactie tussen de laatste twee variabelen als onafhankelijke variabelen.*

	<b>Model 1</b> b/beta (S.E.)	<b>Model 2</b> b/beta (S.E.)	<b>Model 3</b> b/beta (S.E.)	<b>Model 4</b> b/beta (S.E.)
Intercept	7.026*** (.690)	6.485*** (.582)	6.607*** (.545)	5.535*** (.518)
Geslacht	-5.471/-253*** (.718)	-2.770/-128*** (.623)	-1.265/-058* (.599)	-1.868/-086*** (.562)
Immigratieachtergrond	1.314/.061 (.721)	2.448/.113*** (.611)	2.248/.104*** (.572)	2.416/.112*** (.534)
Cohorte	3.088/.143*** (.716)	.780/.036 (.617)	-.647/-030 (.591)	-.203/-009 (.554)
Criminale geneigdheid		5.872/.545*** (.319)	4.708/.437*** (.317)	4.322/.401*** (.298)
Lifestyle risk (Blootstelling aan criminogene settings)			3.551/.330*** (.324)	2.412/.224*** (.320)
Criminale geneigdheid x lifestyle risk (interactieterm)				2.518/.294*** (.227)
<b>R<sup>2</sup></b>	.086	.351	.433	.506
<b>Verandering in F- waarde</b>	26.147***	339.023***	119.768***	123.005***

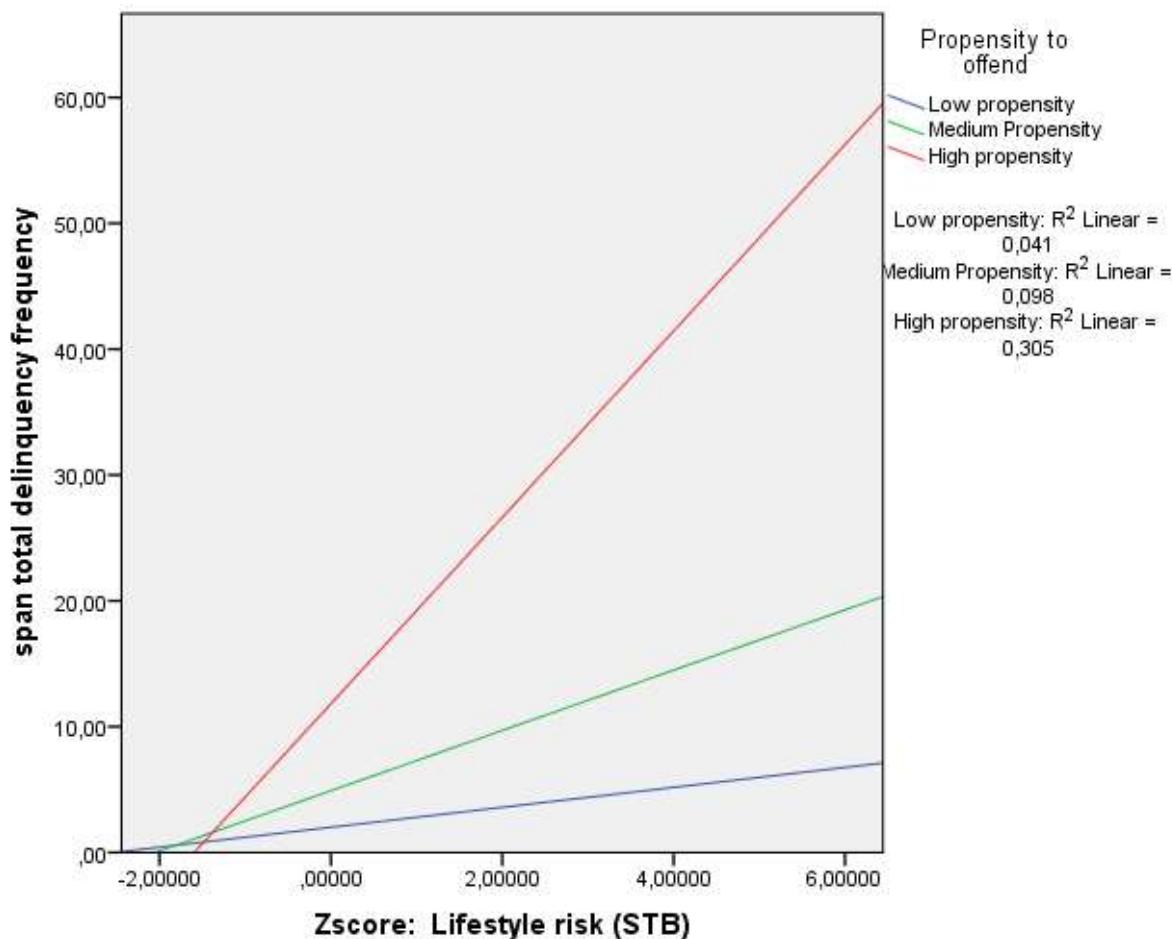
\* p = < .05 ; \*\* = p < .01 ; \*\*\* = p < .001

Tabel 6 toont de resultaten van een bloksgewijze regressieanalyse. In model 1 worden de coëfficiënten van een aantal controlevariabelen gepresenteerd (geslacht, immigratieachtergrond en cohorte waartoe men behoort). In model twee wordt ‘criminale geneigdheid’ toegevoegd aan de analyse. We zien dat criminale geneigdheid een sterk positief effect heeft op zelf gerapporteerde criminaliteit (Beta .545; p<.001). Model 3 voegt de onafhankelijke variabele lifestyle risk (of blootstelling aan criminogene settings) toe. Ook deze variabele heeft een sterk en positief netto-effect op zelf gerapporteerde criminaliteit (Beta ;330; p<.001). Merk op dat het effect van criminale geneigdheid in model 3 lichtjes daalt maar sterk positief en significant blijft (Beta .437; p<.001). Tot slot voegen we in model

4 de interactieterm tussen criminale geneigdheid en lifestyle risk toe. Het interactie-effect is sterk positief en bovendien significant (Beta .294; p<.001). Kijken we naar de determinatiecoëfficiënt dan zien we dat toevoeging van de interactieterm de verklaringswaarde van het model verhoogt met 7.3% ( $R^2$  in model 4= .506 en  $R^2$  in model 3= .433). De verandering in de F-waarde die bij de determinatiecoëfficiënt hoort is eveneens significant, hetgeen wil zeggen dat het model met de interactieterm de data beter samenvat dan het statistische regressiemodel zonder de interactieterm. Alle variabelen in het volledige model verklaren 50.6% van de variantie in zelf gerapporteerde criminaliteit. Bemerk ook nog dat criminale geneigdheid (Beta .401; p<.001) en lifestyle risk (Beta .224; p<.001) wel een positief netto-effect op zelf gerapporteerde criminaliteit behouden. De netto effecten zijn weliswaar minder sterk maar ze blijven wel significant.

We geven het interactie-effect ook grafisch weer in figuur 4.

*Figuur 4: visuele weergave van de interactie tussen criminale geneigdheid (propensity to offend) en lifestyle risk*



Figuur 4 geeft het verband weer tussen blootstelling aan criminogene settings (het aantal uur aanwezig op plaatsen met lage sociale cohesie en delinquenten vrienden) en het aantal delicten per uur (totale som). Dit voorbeeld is afkomstig uit een unieke studie waarbij de kenmerken van de omgeving werden gemeten aan de hand van een bijzondere methode: de STB- space time budget methode. Dit is een soort van dagboek waarin uur per uur de blootstelling van de respondent aan kenmerken in geografische micro-plaatsen wordt voorgesteld.

We zien dat er een heel duidelijke link is tussen lifestyle risk (of het aanwezig zijn op plaatsen met lage sociale cohesie en delinquenten vrienden) en het aantal delicten dat men pleegt, maar we zien ook dat het verband niet voor iedereen hetzelfde is. De regressie-of voorspellingslijnen lopen niet parallel. Voor respondenten met een hoge criminale geneigdheid (high propensity) is de regressierechte veel steiler dan voor de respondenten met medium en lage criminale geneigdheid (medium en low propensity). Voor respondenten met lage criminale geneigdheid is de voorspellingsrechte veel vlakker. De hellingsgraad toont dus de sterkte van het effect van lifestyle risk dat veel sterker is voor respondenten met hoge criminale geneigdheid en zwakker voor respondenten met lage criminale geneigdheid. Concreet betekent dit dat jongeren die laag scoren op het kenmerk ‘geneigdheid tot regelovertraving’ nauwelijks beïnvloed worden door een criminogene omgeving, maar je ziet ook dat onder jongeren die hoog scoren op *propensity to offend* (dit betekent: sterk geneigd zijn regelovertraving als alternatief te zien) een sterke relatie bestaat tussen blootstelling aan criminogene settings en het aantal delicten dat men pleegt per uur in criminogene settings.

#### 4. De pad-analyse

De pad-analyse dankt zijn naam aan de analyse van “paden”, dit zijn statistische paden of manieren waarlangs een variabele een andere beïnvloedt. De klassieke pad-analyse bestaat uit een reeks van structurele vergelijkingsmodellen tussen gemeten variabelen. Een pad-analyse: analyse van structurele modellen waarbij alle variabelen geobserveerd (manifest) zijn. De methode stamt oorspronkelijk uit de biologie en is ontworpen door de bioloog Sewell Wright in 1922. Wanneer de pad-coëfficiënten gestandaardiseerd zijn, kan men ze vergelijken met de gestandaardiseerde regressiegewichten, dus de  $\beta$ -coëfficiënten.

Een structureel model represeneert alle causale hypothesen omtrent de patronen van directe en indirecte effecten tussen alle variabelen in een statistisch model. Dit is een zeer vaak gebruikte manier om hypothesen te toetsen uit criminologische theorieën. **Elk theoretisch construct in het model (bvb. zelf-gerapporteerde criminaliteit, ouderlijke controle,**

**criminele geneigdheid en blootstelling aan criminogene morele settings,...)** wordt gemeten door **1 geobserveerde variabele** en correspondeert dus met **1 variabele in de dataset.**

### ***Directe en indirecte effecten***

Door de opeenvolging van directe effecten kunnen in een pad-analyse op een natuurlijke wijze indirecte effecten worden gemodelleerd (mediatie).

We geven een voorbeeld waarbij de volgorde van de variabelen een rol spelen: x1, y1, y2, y3.

X1 is een onafhankelijke variabele of een **exogene variabele**. Een exogene variabele is een variabele waar geen causale effectrelaties (pijlen) toekomen. Er vertrekken enkel pijlen naar andere variabelen (y1 en y3).

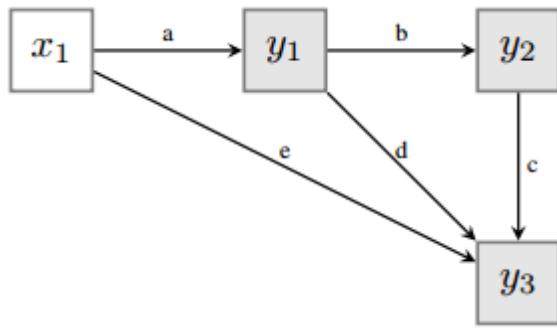
De Y-variabelen (y1, y2 en y3) zijn allemaal **endogene variabelen**. Dit wil zeggen dat er pijlen toekomen. Ze worden causaal bepaald door een reeks andere variabelen.

Y1 wordt rechtstreeks bepaald door x1. Y2 wordt indirect bepaald door x1 en direct door y1. Y3 wordt bepaald door het rechtstreekse effect van x1, het rechtstreekse effect van y1 en het rechtstreekse effect van y2.

*Naast al deze rechtstreekse effecten zijn er nog een reeks onrechtstreekse effecten. Dit zijn mediërende effecten of onrechtstreekse effecten. De onrechtstreekse effecten zijn causale paden waarlangs een variabele een effect uitoefent op een andere variabele, maar nooit rechtstreeks, altijd via de impact op een tussenliggende variabele.*

We geven een voorbeeld door alle indirecte effecten te bespreken die uit de figuur af te leiden zijn.

- Er is een indirect effect vast te stellen van x1 op y3, met name een indirect effect dat verloopt via y1 en y2 naar y3, maar ook een effect via y1 naar y3.
- e = het rechtstreekse effect van x1 op y3
- a = het rechtstreekse effect van x1 op y1
- b = het rechtstreekse effect van y1 op y2
- c = het rechtstreekse effect van y2 op y3
- d = het rechtstreekse effect van y1 op y3



We kunnen dit pad-model ook uitschrijven aan de hand van een reeks vergelijkingsmodellen (equations). Dit doen we door de regressievergelijking uit te schrijven voor elke endogene variabele. Bij het uitschrijven vermelden we enkel de rechtstreekse effecten. De onrechtstreekse effecten zijn af te leiden uit de vergelijkingen. Wanneer we de vergelijkingen tekenen, krijgen we zicht op de mogelijke manieren waarop de variabelen met elkaar in verband staan.

Laat  $x_1$  de sociale controle in het gezin zijn,  $y_1$  de criminale geneigdheid en  $y_2$  de blootstelling aan criminogene morele settings. Sociale controle heeft effecten op zelf-gerapporteerde criminaliteit via de impact op de criminale geneigdheid, maar heeft ook een rechtstreeks effect op de zelf-gerapporteerde criminaliteit. De criminale geneigdheid heeft op diens beurt ook een rechtstreeks effect op de zelf-gerapporteerde criminaliteit, maar ook onrechtstreeks via de blootstelling aan criminogene morele settings.

$$Y_1 = ax_1 + e_1$$

$$Y_2 = by_1 + e_2$$

$$Y_3 = ex_1 + dy_2 + cy_1 + e_3$$

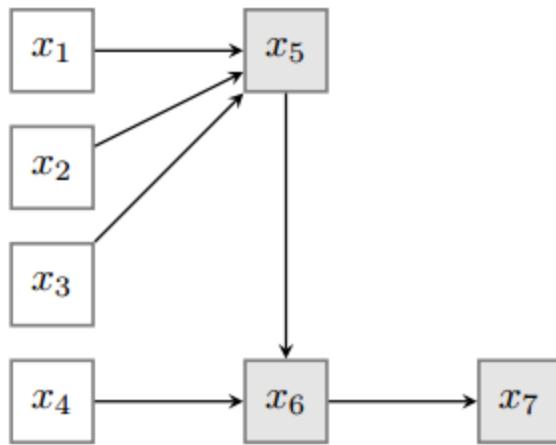
## 5. Berekening van de totale en indirecte effecten in de pad-analyse

**Een totaal effect = alle rechtstreekse effecten + alle onrechtstreekse effecten**

- **Het totale effect van  $x_1$  op  $y_1$**  = de bivariate correlatie ( $x_1, y_1$ ) of de  $\beta$ -coëfficiënt in de regressie van  $y_1$  op  $x_1$ .
- **Het totale effect van  $x_1$  op  $y_2$**  = Het rechtstreekse effect van  $x_1$  op  $y_2$  (= 0 want er is geen rechtstreekse pijl) + het onrechtstreekse effect van  $x_1$  op  $y_2$  via  $y_1$  en dat is gelijk aan  $a \cdot b$ .

- Het totale effect van  $x_1$  op  $y_3$  = de som van het rechtstreekse effect van  $x_1$  op  $y_3$  en de diverse onrechtstreekse effecten. Hier hebben we een direct effect (e), een binnenste onrechtstreeks effect (van  $x_1$  op  $y_3$  via  $y_1$ ) ( $a^*d$ ) en een buitenste onrechtstreeks effect (van  $x_1$  op  $y_3$  via  $y_1$  en  $y_2$ ) ( $a^*b^*c$ ). Dus de berekening van het totale effect =  $e + a^*d + a^*b^*c$ . Nog steeds geldt dat er geen onrechtstreeks effect is van  $x_1$  op  $y_3$  via  $y_2$  omdat er geen rechtstreeks effect is van  $x_1$  op  $y_2$ .

## 6. Nog een voorbeeld van een pad-model



### Overzicht

In deze pad-analyse zijn er meerdere exogene variabelen. De exogene variabelen zijn  $x_1 - x_4$ , want hier komen geen pijlen toe.

Er zijn meerdere endogene variabelen. Een aantal endogene variabelen heeft een dubbel statuut:  $x_5-x_6$  = mediatoren = ze mediëren de effecten van de onafhankelijke variabelen.

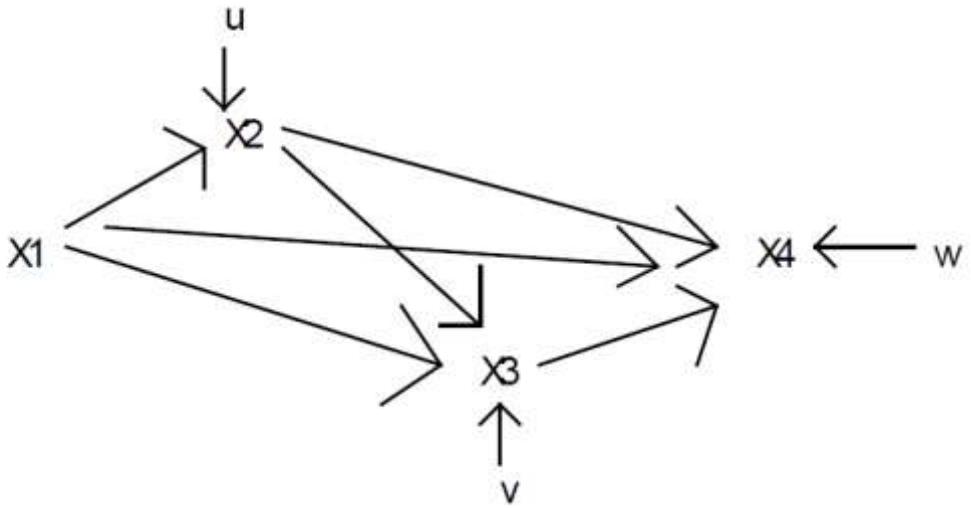
$x_7$ : uitkomstvariabele of hoofdvariabele.

We kunnen  $x_1 \rightarrow x_7$  met een omweg bereiken.

Ook dit pad-model is een voorstelling van de realiteit: het is de hypothese van de onderzoeker, die zich door een theorie heeft laten inspireren.

Laten we de rekenoefening eens maken aan de hand van een intussen klassiek geworden voorbeeld uit de sociologie, uit de cursus van Otis Dudley Duncan, één van de klassieke handboeken over pad-analyse.

Ziehier een reeks van causale effecten.



Laten we dit pad-model eens onder de loep nemen en elk effect bespreken. Om de paden te identificeren kunnen we twee manieren volgen: we werken van de oorzaken naar de gevolgen of we werken omgekeerd, van de gevolgen naar de oorzaken. Het hierboven getekend pad-model is een recursief model, dit wil zeggen dat er geen feedback loops zijn, of geen wederzijdse causale effecten.

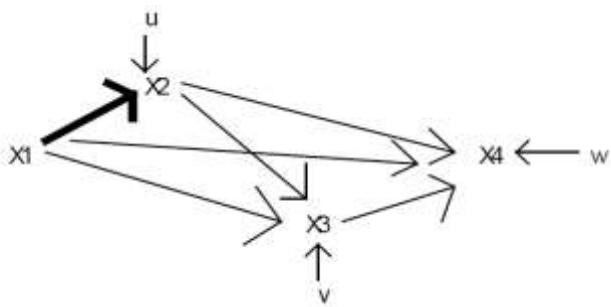
Er is een direct effect van  $x_1$  op  $x_2$ .

Wanneer we paden bestuderen, dan geldt de volgende regel:

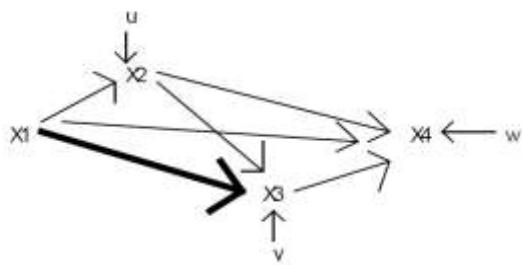
Ga na hoe je van de variabelen aan de linkerkant naar de variabelen aan de rechterkant geraakt. Elke pijl wijst op een route.

- Als je één variabele kan terugkeren en dan kan stoppen, heb je een direct effect geïdentificeerd.
- Als je twee variabelen kan terugkeren en niet opnieuw terug vooruit kan, heb je een exclusief indirect effect geïdentificeerd.
- Als je een variabele kan vinden die een pijl trekt naar twee variabelen die verderop in het model staan, heb je een gemeenschappelijke oorzaak ontdekt.

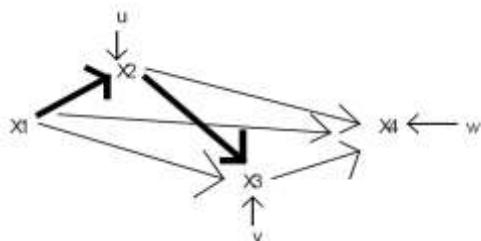
We leggen het principe van de pad-analyse uit aan de hand van een voorbeeld waarbij we wijzen op de verschillende effecten tussen  $x_1$  en  $x_4$ .



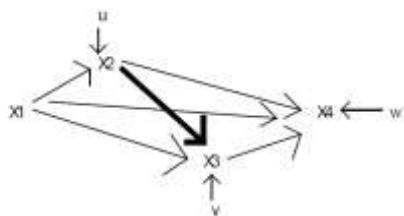
In de figuur hierboven werd het directe effect van  $x_1$  naar  $x_2$  in de verf gezet. De variabele “ $u$ ” is de aliénatiecoëfficiënt, ofwel de proportie onverklaarde variantie in  $x_2$  op basis van  $x_1$ .



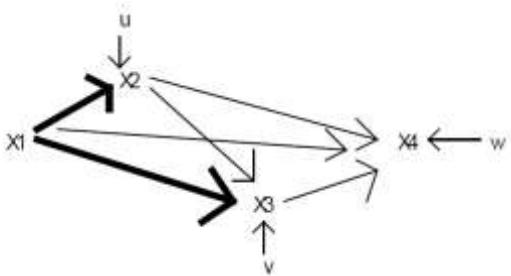
Vervolgens zie je dat er ook een rechtstreeks effect is van  $x_1$  op  $x_3$ . Ook dat is een rechtstreeks effect.



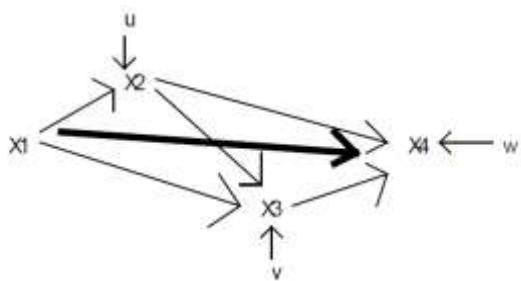
Hierboven zie je dat er niet enkel een rechtstreeks effect van  $x_1$  naar  $x_3$  bestaat, maar dat je ook van  $x_1$  naar  $x_3$  geraakt via  $x_2$ . Dit is een indirect effect.



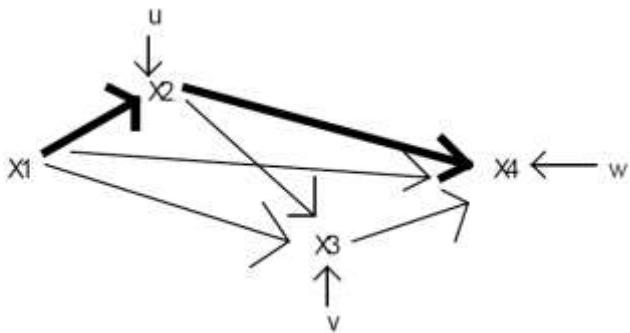
Het effect van  $x_2$  op  $x_3$  is ook een rechtstreeks effect.  $x_3$  kan verklaard worden vanuit de rechtstreekse effecten van  $x_1$  en  $x_2$ . De variabele “ $v$ ” verwijst hier naar de onverklaarde variantie in  $x_3$ , die niet door  $x_1$  en  $x_2$  kan worden verklaard.



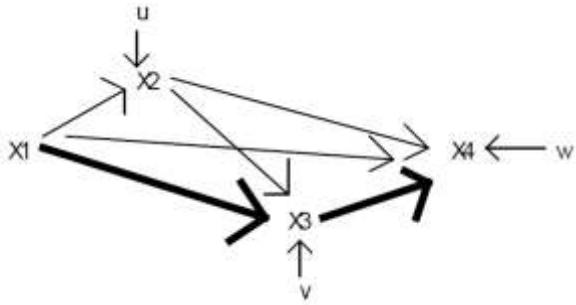
We zien echter ook dat  $x_1$  kan gelden als een partiële gemeenschappelijke oorzaak van de relatie tussen  $x_2$  en  $x_3$ . We zeggen partiële gemeenschappelijke oorzaak, omdat er ook nog een rechtstreeks effect bestaat van  $x_2$  op  $x_3$ . Indien er geen pijl kon getrokken worden van  $x_2$  naar  $x_3$ , dan zou de samenhang tussen  $x_2$  en  $x_3$  volledig te wijten zijn aan het feit dat  $x_2$  en  $x_3$  een gemeenschappelijke oorzaak kennen.



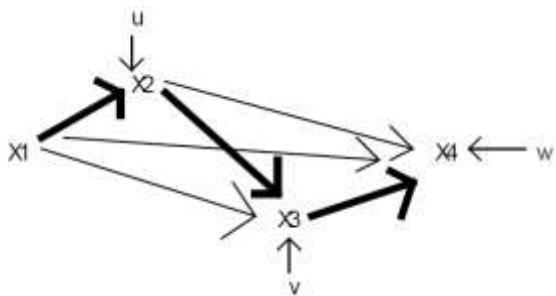
Verder is er ook een effect dat rechtstreeks gaat van  $x_1$  naar  $x_4$ . De variabele “ $w$ ” is de niet verklaarde variantie in  $x_4$ . Dit is de variantie in  $x_4$  die niet kan worden verklaard door de rechtstreekse effecten van  $x_1$ ,  $x_2$  en  $x_3$ .



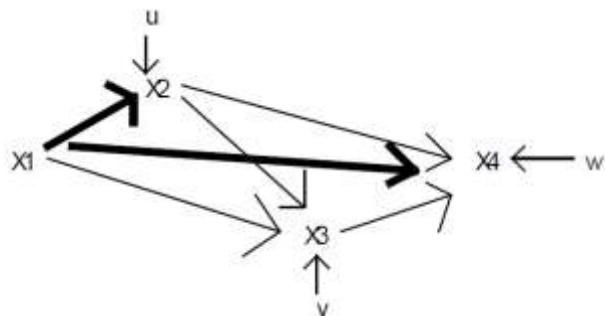
Er is naast het rechtstreeks effect van  $x_1$  op  $x_4$  ook een onrechtstreeks effect van  $x_1$  op  $x_4$  via  $x_2$ .



Er is naast het rechtstreeks effect van  $x_1$  op  $x_4$  ook een onrechtstreeks effect van  $x_1$  op  $x_4$  via  $x_3$ .

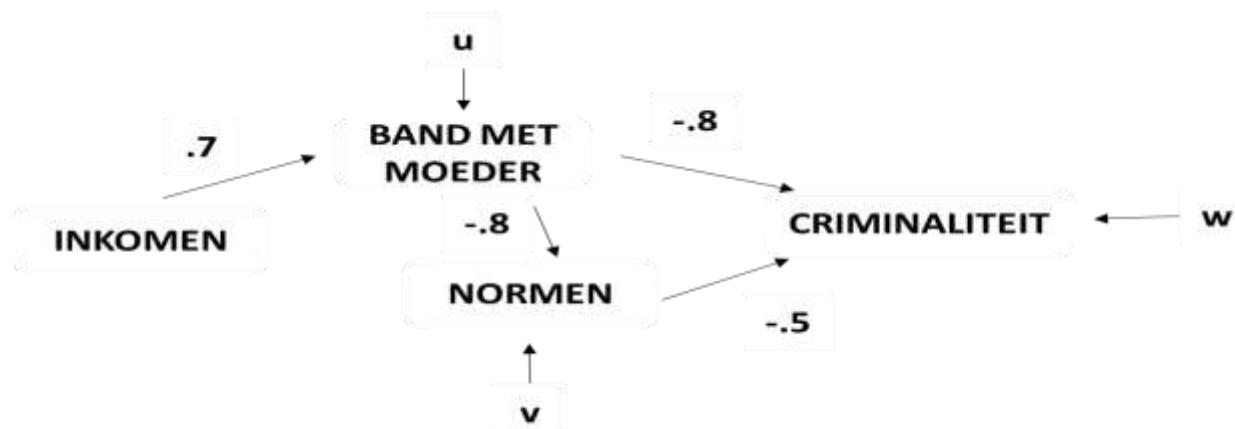


Verder is er nog een manier om van  $x_1$  naar  $x_4$  te geraken: namelijk via  $x_2$  naar  $x_3$  en zo naar  $x_4$ .



En tenslotte is er ook een gemeenschappelijke oorzaak vast te stellen: de relatie tussen  $x_2$  en  $x_4$  heeft  $x_1$  als gemeenschappelijke oorzaak. Het verband is echter niet volledig te wijten aan het fenomeen van de gemeenschappelijke oorzaak, want er is nog het overblijvende rechtstreekse effect van  $x_2$  op  $x_4$ .

## 7. Een (fictief) rekenvoorbeeld op basis van de gestandaardiseerde padcoëfficiënten

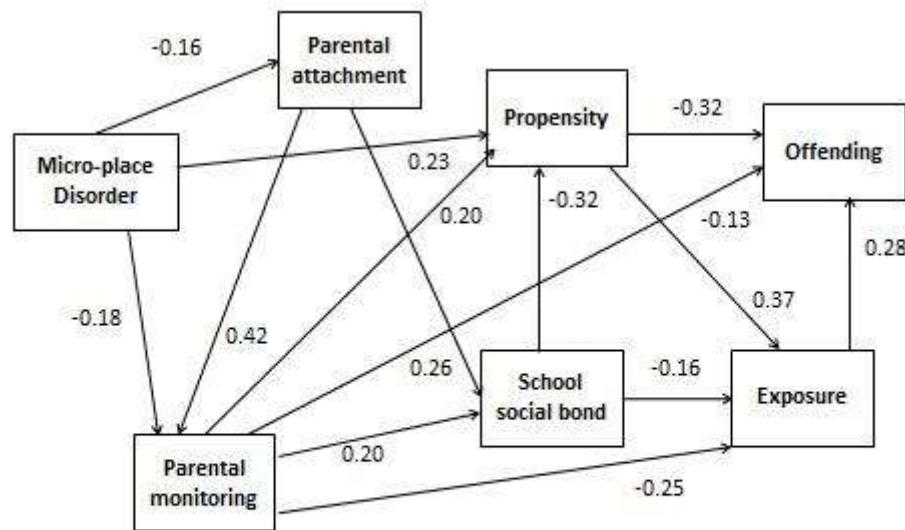


Aan de hand van de methode van Sewell-Wright kan de bivariate correlatie tussen twee variabelen gevonden worden op basis van de gestandaardiseerde padcoëfficiënten.

Correlatiecoëfficiënt	Sewell-Wright methode
R BAND MOEDER; Inkomen = 0.7	Dit is de correlatie tussen band met de moeder en inkomen. <i>Aangezien er enkel een rechtstreeks effect is, is het berekenen van de correlatiecoëfficiënt eenvoudig: deze is gelijk aan de gestandaardiseerde padcoëfficiënt.</i> Dit is het rechtstreeks effect van inkomen op band met de moeder.
R NORMEN; BAND MOEDER = -0.8	Rechtstreeks effect van band met de moeder op normen.
R NORMEN; Inkomen = $-0.8 * 0.7 = -0.56$	Dit is het indirecte effect van inkomen – inkomen beïnvloedt band met de moeder, dat zelf van invloed is op normen.
R CRIMINALITEIT; NORMEN = $-0.5 + -0.8 * -0.8 = 0.14$	De correlatie tussen criminaliteit en normen is gelijk aan <i>het rechtstreekse effect plus de gemeenschappelijke oorzaak</i> . Band met de moeder is een gemeenschappelijke oorzaak van zowel normen als criminaliteit.
R CRIMINALITEIT; BAND MOEDER = $-0.8 + -0.8 * -0.5 = -0.4$	De correlatie tussen criminaliteit en band met moeder is gelijk aan <i>het rechtstreekse effect plus het onrechtstreekse effect</i> .
R CRIMINALITEIT; Inkomen = $-0.8 * 0.7 * -0.5 * -0.8 * 0.7 = -0.28$	Inkomen kent een <b>EERSTE</b> indirect effect op criminaliteit – inkomen is van invloed op band met de moeder, dat op zich van invloed is op criminaliteit.

	<p>Maar er is nog een <b>TWEDE</b> indirect pad vertrekend vanuit inkomen. Inkomen is van invloed op band met de moeder, dat op zich van invloed is op normen, dat van invloed is op criminaliteit.</p>
--	---

### Een Belgisch criminologisch voorbeeld: jeugddelinquentie vanuit een geïntegreerde theorie



Antwerp Early Adolescence School Survey (2005)

De pad-analyse die we hierboven presenteren komt uit een reëel gevoerd onderzoek over de relatie tussen onrechtstreekse oorzaken van regelovertredend gedrag en rechtstreekse oorzaken van regelovertredend gedrag. Het gaat om een analyse van gegevens afkomstig uit een grootschalige studie naar jeugddelinquentie en de rol van de woonomgeving. Het onderzoek eindigde met de vaststelling dat buurten niet zo belangrijk waren als oorspronkelijk werd aangenomen, maar dat de micro-context onrechtstreeks wel relevant is. De micro-context is de plaats zelf (de straat of het straatsegment) en niet de volledige buurt waar men woont. Er werden enkele veronderstellingen gemaakt over de mechanismen die belangrijk zijn bij de verklaring van individuele verschillen in delinquent gedrag en de onderlinge relatie tussen directe en indirecte invloedssferen.

De geïntegreerde theorie (“conditions-controls-exposure theory” van Pauwels en Hardyns) gaat ervan uit dat de micro-context indirect een invloed heeft op regelovertredend gedrag, via

de aantasting van sociale bindingen (attachment) en ouderlijk toezicht (monitoring), maar ook via de invloed op de geneigdheid om regelovertraving als alternatief te zien (propensity). Dat idee lijkt goed stand te houden want je ziet in de figuur, die vergezeld is van gestandaardiseerde pad-coëfficiënten enkel indirecte effecten van de concentratie van overlast in de micro-plaats waar men woont op delinquentie. Er zijn rechtstreekse effecten op attachment, monitoring en propensity. Verder vallen nog een aantal dingen op:

- Er zijn rechtstreekse effecten van parental attachment op parental monitoring en de sociale band met de school. Deze bevindingen kunnen worden geïnterpreteerd vanuit de bindingentheorie. In tegenstelling tot de klassieke analyse van Hirschi, leren we met deze padanalyse veel meer, met name dat er toch verschillende onderlinge relaties zijn tussen de elementen van de sociale band.
- Er zijn een aantal rechtstreekse effecten van monitoring: een rechtstreeks effect op de sociale band met de school, een rechtstreeks effect op propensity en een rechtsreeks effect op blootstelling aan criminogene situaties (de ongestructureerde leefstijl).
- Er zijn slechts twee directe effecten van de sociale band met de school en dat is een effect op propensity en exposure.
- Propensity heeft het sterkste rechtstreekse effect op zelfgerapporteerde criminaliteit en heeft ook een sterk effect op de mate van blootstelling aan criminogene settings.
- Blootstelling aan criminogene settings heeft een rechtsreeks effect op regelovertrijdend gedrag.

Pad-analyses leveren dus een schat aan informatie op voor zowel de evaluatie van criminologische theorieën als voor het beleid.

## 8. Leerdoelen

In dit hoofdstuk werd een aanzet gegeven tot het bestuderen van meer complexe relaties tussen variabelen. De techniek die daarvoor gebruikt wordt is de padanalyse. Deze techniek kent zijn oorsprong in de biologie en het was de bioloog Sewell-Wright die een methode bedacht om op basis van de pad-coëfficiënten, die niet meer zijn dan gestandaardiseerde richtingscoëfficiënten, terug te keren naar de correlatiecoëfficiënten. We verwachten van studenten dat ze, als hen een pad-model wordt getoond, de rechtstreekse, onrechtstreekse en totale effecten kunnen berekenen. We verwachten eveneens dat uit een pad-diagram de bivariate correlaties kunnen berekend worden tussen een aantal gevraagde variabelen. Maar

bovenal is het belangrijk dergelijke analyses te kunnen begrijpen wanneer je ze tegenkomt in de kwantitatieve criminologische literatuur.



## **Slotbeschouwingen**

Hiermee zijn we aan het einde van dit inleidende handboek in de toegepaste beschrijvende en inferentiële statistiek voor criminologen gekomen. We hopen aan het einde van dit handboek dat volgende zaken duidelijk zijn: (a) statistiek is een hulpmiddel voor de criminoloog en dient in de eerste plaats om beschrijvende, verkennende en toetsende onderzoeks vragen en probleemstellingen te beantwoorden; (b) een niet correct gebruik van deze technieken kan leiden tot een foutieve voorstelling van de empirische gegevens; (c) een kritische houding is nodig om het naakte cijfer van een correcte interpretatie te voorzien.

Met dit handboek hopen we te hebben aangetoond wat er achter de schermen gebeurt wanneer onderzoekers grootschalige gegevensbestanden analyseren aan de hand van statistische software. Zelf zijn we een enorme voorstander van het gebruik van software in de sociale wetenschappen, want deze reduceert het aantal fouten bij het berekenen van verbanden, uiteraard mits correcte toepassing van de statistische regels, de keuze voor de juiste analysetechniek en dergelijke meer. We hopen dat we de persisterende kwantifobie of angst voor statistiek onder studenten hebben kunnen wegnemen door de mathematische achtergronden te beperken en te hebben voorzien van concrete criminologische toepassingen. Door deze klemtoonverschuiving kan meer tijd worden besteed aan de interpretatie van bevindingen uit reëel criminologisch onderzoek.



# **OPLOSSINGEN TESTVRAGEN**



## OPLOSSINGEN TESTVRAGEN

### UNIVARIATE STATISTIEK

#### 1. Deviatiescores zijn

- De som van de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde
- De som van de afwijkingen tegenover het rekenkundig gemiddelde

#### 2. De mediaan is een robuuste parameter van centraliteit

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

#### 3. De Index van Diversiteit kan geïnterpreteerd worden als de kans dat twee willekeurig gekozen onderzoekseenheden tot een verschillende categorie behoren

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

#### 4. De variatieratio neemt de waarde van nul aan indien alle waarnemingen tot de modale categorie behoren

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

#### 5. Een variabele van het metrisch niveau kan bestudeerd worden op ordinaal niveau

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

#### 6. Een variabele van het ordinale niveau kan bestudeerd worden op het metrische niveau

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**7. De variatiebreedte is het verschil tussen de maximale waarde en de minimale waarde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**8. De variatiebreedte is de som van de maximale en minimale waarde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**9. Uitspraken extrapoleren van de steekproef naar de populatie doe je via**

- de beschrijvende univariate statistiek
- de inferentiële statistiek

**10. De steekproefstandaardafwijking wordt berekend op basis van de formule van de populatiestandaardafwijking maar in de noemer staat  $n+1$**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**11. Een bar chart wordt gebruikt**

- Voor variabelen op het metrische niveau
- Voor variabelen op het categorische niveau

**12. Variabelen die op een histogram worden gepresenteerd zijn steeds ratio niveau**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**13. Een lijndiagram wordt gebruikt voor de**

- Visuele voorstelling van metrische gegevens die niet in klassen zijn gegroepeerd (ruwe scores)
- Visuele voorstelling van metrische gegevens die wel in klassen zijn gegroepeerd

**14. Een frequentiepolygoon wordt gebruikt voor de**

- Visuele voorstelling van metrische gegevens die niet in klassen zijn gegroepeerd (ruwe scores)
- Visuele voorstelling van metrische gegevens die wel in klassen zijn gegroepeerd

**15. Een platykurtische variabele is**

- Platter dan een normaal verdeelde variabele
- Scherper dan een normaal verdeelde variabele

**16. Een box-plot kan worden gebruikt voor variabelen vanaf**

- Het nominaal niveau
- Ordinaal niveau
- Interval niveau
- Ratio niveau

**17. De mediaan komt overeen met**

- Het vijftigste percentiel
- Het eenenvijftigste percentiel

**18. Als we een kenmerk dat perfect normaal verdeeld is voorstellen via een box-plot, dan is de afstand tussen de mediaan en de hoogste waarde even groot als de afstand tussen de mediaan en de laagste waarde**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**19. Een variabele die rechtsscheef verdeeld is**

- Heeft een langere staart naar rechts
- Heeft een langere staart naar links
- Heeft geen staart

**20. De interkwartielafstand is een maat van**

- Centraliteit
- **Spreiding**
- Vorm

**21. Een frequentieverdeling kunnen we opvatten als een kansverdeling**

- **Deze uitspraak is juist**
- Deze uitspraak is verkeerd

**22. Een onderzoekseenheid heeft een z-score van -2.20 voor het metrisch kenmerk “studieresultaat” .**

- **De onderzoekseenheid valt buiten de centrale 95% van de waarnemingen.**
- De onderzoekseenheid doet het beduidend beter dan de gemiddelde onderzoekseenheid

**23. Centreren wil zeggen dat men een kenmerk uitdrukt als een afwijking tegenover het rekenkundig gemiddelde**

- **Deze uitspraak is juist**
- Deze uitspraak is verkeerd

**24. De frequentieverdeling (histogram) verandert vormelijk niet wanneer men standaardiseert**

- **Deze uitspraak is juist**
- Deze uitspraak is verkeerd

**25. Het rekenkundig gemiddelde is een spreidingsmaat die gevoelig is voor uitschieters**

- Deze uitspraak is juist
- **Deze uitspraak is verkeerd**

**26. Variabelen van het categorische niveau bevatten categorieën. Een onderzoeksseenheid mag tegelijkertijd in twee categorieën van dezelfde variabele zitten**

- Dit mag niet als we de regels van de statistiek volgen
- Dit mag wel, er zijn hier geen regels voor

**27. Operationalisering betekent**

- Dat we een kenmerk meetbaar maken
- Dat we een kenmerk van een conceptuele definitie voorzien

**28. De populatievariantie is de som van de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde, gedeeld door het steekproefeffectief**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**29. De variatieratio is een spreidingsmaat die we moeten gebruiken wanneer twee metrische kenmerken in een verschillende eenheid werden gemeten en men wil de spreiding vergelijken**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**30. De keuze voor een univariate parameter wordt ingegeven door**

- De onderzoeksvergadering
- Het meetniveau
- Beide

**31. Stel: je bent activist die ijvert voor meer inkomensgelijkheid. Om de inkomenongelijkheid te demonstreren, kan je kiezen tussen een aantal parameters. Inkomen is zeer scheef verdeeld.**

- Je presenteert vanuit je standpunt de mediaan
- Je presenteert vanuit je standpunt het rekenkundig gemiddelde

## BIVARIATE EN INFERENTIEËLE STATISTIEK

**1. De Y-as wordt ook wel het ordinaat genoemd**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**2. De X-as wordt ook wel de abscis genoemd**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**3. Een scatterplot kan gebruikt worden om de relatie tussen een ratio-variabele en een interval-variabele te presenteren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**4. Een scatterplot kan gebruikt worden om de relatie tussen een ordinale variabele en een interval-variabele te presenteren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**5. Een spreidingsdiagram kan gebruikt worden om de relatie tussen een nominale en een ordinale variabele te presenteren**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**6. Het  $(x,y)$  coördinaat van het rekenkundig gemiddelde van  $x$  en het rekenkundig gemiddelde van  $y$  is het bivariate zwaartepunt van het spreidingsdiagram**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**7. De ongestandaardiseerde regressiecoëfficiënt ( $b_1$ ) uit een bivariate regressieanalyse is een**

- Symmetrische maat
- Asymmetrische maat

**8. De bivariate gestandaardiseerde regressiecoëfficiënt uit de regressie van Y op X is gelijk aan de gestandaardiseerde covariantie tussen x en y**

- Deze uitspraak is juist
- Deze uitspraak is fout

**9. De regressie van Q op S wil zeggen dat we**

- Q als afhankelijke variabele hebben en S als onafhankelijke variabele
- S als afhankelijke variabele hebben en Q als onafhankelijke variabele

**10. Het percentageverschil is een symmetrische associatiemaat**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**11. Hierna vind je een tabel waarbij de relatie tussen twee nominale variabelen, X en Y wordt voorgesteld. We veronderstellen dat R de afhankelijke variabele is en dat S een causale invloed uitoefent op R.**

	Variabele S		Totaal
Variabele R	A	B	E
	C	D	F
Totaal	G	H	I

Welke van de uitspraken is juist:

- E en F noemen we kolommarginalen
- E en F noemen we rijmarginalen
- A / G geeft de proportie van A
- A / C geeft de odds op A

**12. Chi-kwadraat is een associatiemaat die zeer gevoelig is aan N**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**13. Als je N vermenigvuldigt met 2, dan wordt de waarde van chi-kwadraat twee keer zo groot**

- Deze uitspraak is juist
- Deze uitspraak is fout

**14. Chi-kwadraat moet worden berekend op basis van de ruwe scores**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**15. Chi-kwadraat kan ook worden berekend op basis van de proporties**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**16. In een 2\*2 tabel is de waarde van Phi gelijk aan de waarde van V**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**17. Bekijk volgende tabel**

Slachtofferschap vijf jaar	afgelopen			Scholingsgraad		
				Laag	Midden	Hoog
Nooit	A	B	C			
Een maal	D	E	F			
Twee maal of meer	G	H	I			

**Welke uitspraken over deze tabel zijn juist**

- ( $E^*F$ ) is een consistent paar
- ( $F^*H$ ) is een inconsistent paar

- Gamma is gebaseerd op de verhouding tussen de consistentie paren min de inconsistentie paren en de consistentie paren plus de inconsistentie paren
- Gamma is een associatiemaat die monotoniciteit veronderstelt
- Gamma is een asymmetrische maat
- Gamma wordt gebruikt op nominaal niveau

#### 18. Welke uitspraken zijn juist?

- Spearman's Rho is een rangcorrelatiecoëfficiënt
- Pearson's r is een rangcorrelatiecoëfficiënt
- Kendall's Tau- is een rangcorrelatiecoëfficiënt
- Rangcorrelaties zijn asymmetrische maten

#### 19. Welke uitspraken over Pearson's r zijn juist?

- De bivariate correlatiecoëfficiënt van Pearson veronderstelt lineariteit (zoniet kan Pearson's r leiden tot verkeerde conclusies)
- De bivariate correlatiecoëfficiënt is een gestandaardiseerde covariantie
- De bivariate correlatiecoëfficiënt wordt berekend uit niet gekwadrateerde deviatiescores

#### 20. De bivariate correlatiecoëfficiënt tussen ouderlijk toezicht en criminaliteit en y is negatief en de gestandaardiseerde rico uit de regressie van criminaliteit op toezicht bedraagt -0.35.

- Daaruit volgt dat toezicht een remmend effect heeft op criminaliteit
- Daaruit volgt dat toezicht nefast is voor de criminaliteit want meer toezicht, resulteert in meer criminaliteit
- Daaruit volgt dat meer toezicht ongerelateerd is aan criminaliteit

#### 21. De covariatie is

- De kruisproductensom
- De som van de deviatiescores van x + de deviatiescores van y

**22. Hieronder volgen een aantal uitspraken over de lineaire regressieanalyse**

- De lineaire regressie is een asymmetrische techniek
- De lineaire regressie (het basismodel) is gebaseerd op de kleinste kwadratenoplossing
- De lineaire regressie is niet geschikt voor kwadratische curvilineaire relaties

**23. Het residu is**

- Het verschil tussen geobserveerde en verwachte waarde
- Het verschil tussen verwacht en geobserveerde waarde

**24. De som van de gekwadrateerde residuelen bedraagt nul**

- Deze uitspraak is juist
- Deze uitspraak is fout

**25. De variatie in Y kan ook uitgedrukt worden als de som van de regressie sum of square minus de residual sum of square**

- Deze uitspraak is juist
- Deze uitspraak is fout

**26. Als de punten uit een puntenwolk heel dicht bij de regressierechte liggen, dan kunnen we vermoeden dat**

- De model fit zeer hoog zal zijn
- De model fit zeer laag zal zijn

**27. Het is perfect mogelijk dat twee regressiecoëfficiënten eenzelfde beta-waarde hebben, maar een verschillende aliénatiecoëfficiënt hebben**

- Deze uitspraak is juist
- Deze uitspraak is fout

**28. De variantieanalyse wordt gebruikt wanneer**

- De onafhankelijke variabele nominaal is en de afhankelijke variabele metrisch
- De afhankelijke variabele nominaal is en de onafhankelijke variabele metrisch

**29. Eta-kwadraat komt qua interpretatie overeen met een determinatiecoëfficiënt**

- Deze uitspraak is juist
- Deze uitspraak is fout

**30. De toets van de significantie van de determinatiecoëfficiënt gebeurt aan de hand van**

- De F-toets
- De T-toets

**31. De toets van het intercept en de rico gebeurt aan de hand van de**

- De F-toets
- De T-toets
- De Z-toets

**32. Intervalschatting is een centraal element uit de**

- Inferentiële statistiek
- Descriptieve statistiek

**33. De steekproevenverdeling van X is**

- De verdeling van X
- De verdeling van alle gemiddelde waarden voor X in een reeks van steekproeven met gelijke omvang
- De verdeling van alle gemiddelde waarden voor X in een reeks van steekproeven met ongelijke omvang
- Geen van de voorgaande beweringen is juist

**34. Een betrouwbaarheidsinterval van met een alfa (kans op vergissing) van 10 %**

= Z-score van 90

- Komt overeen met een z-score van 1.645
- Komt overeen met een z-score van 1.960
- Komt overeen met een z-score van 2.576

**35. Ik doe criminologisch onderzoek naar belastingontduiking en stel vast dat het gemiddeld aantal veroordelingen onder belastingontduikers (mean = 3) lager is dan onder inbrekers (mean = 5). De betrouwbaarheidsintervallen van de beide gemiddeldes blijken elkaar te overlappen.**

Hieruit besluit ik dat

- De gemiddeldes niet significant van elkaar verschillen
- De gemiddeldes wel significant van elkaar verschillen

**36. Een type I fout betekent :**

- De nulhypothese is juist en ik verworp ze foutief
- De nulhypothese is verkeerd en ik behoud ze foutief

**37. Een type II fout betekent :**

- De nulhypothese is juist en ik verworp ze foutief
- De nulhypothese is verkeerd en ik behoud ze foutief

**38. De power van een test berekenen is belangrijk want hierdoor houdt men rekening met zowel type I als type II fouten**

- Deze uitspraak is juist
- Deze uitspraak is fout

**39. De tussengroepsvariantie is de tussengroepsvariatie gedeeld door het aantal vrijheidsgraden tussen groepen**

- Deze uitspraak is juist
- Deze uitspraak is fout

**40. Als de tussengroepsvariantie groter is dan de binnengroepsvariantie, en de F-toets geeft een waarde die hoger is dan de kritische waarde, is het verband tussen x en y significant**

- Deze uitspraak is juist
- Deze uitspraak is fout

**41. Als de tussengroepsvariantie groter is dan de binnengroepsvariantie, en de F-toets geeft een waarde die lager is dan de kritische waarde, is het verband tussen x en y significant**

- Deze uitspraak is juist
- Deze uitspraak is fout

## MULTIVARIATE STATISTIEK

**1. De meervoudige regressieanalyse met X1 en X2 als onafhankelijke variabelen geeft hetzelfde resultaat als twee afzonderlijke bivariate regressieanalyses. Klopt dat?**

- Dit klopt enkel in de situatie waarbij  $r(x_1, x_2) = 0$ .
- Dit klopt nooit

**2. R (hoofdletter) staat in de output van een multipele regressie voor de correlatie tussen Y en de verwachte waarde voor Y op basis van de onafhankelijke variabelen**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**3. Een meervoudige multivariate regressieanalyse is een**

- Regressieanalyse met meerdere onafhankelijke variabelen en meerdere afhankelijke variabelen
- Regressieanalyse met meerdere onafhankelijke variabelen en slechts een afhankelijke variabele
- Regressieanalyse met meerdere afhankelijke variabele en een onafhankelijke variabele

**4. Heteroscedascitieit wil zeggen dat**

- De waarde van de residuele termen toenemen naarmate X1 toeneemt
- De waarde van de residuele termen gelijk blijft, naarmate X1 toeneemt

**5. Additiviteit betekent dat X1 en x2 onafhankelijke effecten hebben, dit wil zeggen: ze dragen elk bij tot de verklaring van Y**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**6. Lineariteit wil zeggen dat Y als lineaire functie van X kan worden uitgedrukt**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**7. Curvilineariteit wil zeggen dat het effect van X1 op Y toeneemt of afneemt naargelang X1 toeneemt.**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**8. Een uitbijter kan de regressielijn ernstig beïnvloeden**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**9. Interactie betekent dat het effect van X1 op Y conditioneel is op X2**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**10. Een mediatorvariabele is een variabele die het effect van een exogene variabele op een afhankelijke variabele medieert (dwz dat deze variabele het effect van de exogene variabele wegverklaart en dat het effect van de exogene variabele op Y via de mediator verloopt)**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

**11. Een totaal effect is in een pad-analyse gelijk aan de rechtstreekse effecten min de onrechtstreekse effecten**

- Deze uitspraak is juist
- Deze uitspraak is verkeerd



# **SYNTHESE OEFENING**



## **LEES EERST DIT**

Hierna vind je een synthese oefening. Dit is een ultieme rekentest, waarbij je uiterste concentratie wordt gevraagd.

**Het databestand van deze oefening is het format waarin de examenvragen zullen gepresenteerd worden.**

**Het is belangrijk dat je je hiermee vertrouwd maakt.**

De rijen in het databestand vertegenwoordigen statistische eenheden, de kolommen vertegenwoordigen de variabelen. Elke cel geeft voor één statistische eenheid de waarde weer op de betreffende variabele.

Op het examen worden meerdere vragen gesteld op basis van een analyse met gegevens afkomstig uit één databestand (in het format zoals hier in deze synthese oefening gepresenteerd). We testen je kennis over de univariate en bivariate statistiek dan op basis van één dataset, in plaats van op basis van verschillende datasets. Dat geeft minder rekenwerk. Let wel op dat fouten kunnen accumuleren, maar dat is altijd zo wanneer statistische parameters moeten worden berekend die samengesteld zijn uit verschillende tussenstappen.

We weten uit ervaring dat studenten moeite hebben met deze synthese oefening omdat alle leerstof samengebracht wordt. Studenten hebben de neiging om de leerstof in te studeren ‘in aparte hoofdstukken’. Wanneer alles gesynthetiseerd wordt in één grote oefening, dan loopt het vaak mis. Vandaar dat we ten zeerste aanraden om deze synthese oefening te maken als voorbereiding op het examen.

De oplossingen worden hier nog niet meegegeven en dit om didactische redenen, maar worden in de loop van de lessenreeks ter beschikking gesteld via Ufora. Meer informatie hieromtrent wordt later meegedeeld.

**VEEL SUCCES!!**

## SYNTHESE-OEFENING

Sommige wetenschappers zijn van mening dat de kwaliteit van de relatie tussen adolescent en ouders een belangrijke factor in de relatie tot van huis weglopen. Op basis van een kleine toevalsstekproef van adolescenten werden de volgende data verzameld met betrekking tot sexe, het aantal uur dat respondenten wekelijks kwaliteitstijd doorbrengen met hun ouders (variabele X1), het aantal keren dat zij in de afgelopen zes maand overwogen om thuis weg te lopen (variabele X2), en studieresultaten (variabele Y).

Adolescent	Sexe	X1 Uren kwaliteitstijd met de ouders	X2 Aantal keren overwogen om thuis weg te lopen	Y Studieresultaten
1	Meisje	1	10	5
2	Meisje	10	2	13
3	Meisje	5	7	7
4	Meisje	7	1	16
5	Meisje	20	0	11
6	Meisje	12	2	17
7	Meisje	5	1	15
8	Meisje	19	0	9
9	Jongen	4	6	7
10	Jongen	10	1	16
11	Jongen	12	1	13
12	Jongen	4	0	5
13	Jongen	10	0	12
14	Jongen	6	1	9
15	Jongen	2	9	5
16	Jongen	9	1	11

- Bereken de gemiddelde waarde voor de variabelen X1, X2 en Y.
- Wat is het gemiddelde van de jongens op X1, X2 en Y?
- Wat is het gemiddelde van de meisjes op X1, X2 en Y?
- Verschillen jongens en meisjes statistisch significant op X1, X2 en Y?
- Wat is de variatie, variantie, standaardafwijking voor X1, X2 en Y?
- Bereken de variatie, variantie, standaardafwijking afzonderlijk voor de jongens en de meisjes
- Wat is de covariatie, covariantie en correlatie tussen X1 en Y, tussen X2 en Y, tussen X1 en X2?
- Zijn deze bivariate associatiematen gelijk voor de jongens als voor de meisjes?
- Bereken de parameters van de regressieanalyse voor de regressie van Y op X1, en voor de regressie van X1 op Y.
- Bereken de correlatie tussen de verwachte waarde voor Y op basis van X1 en de geobserveerde waarde voor Y. Is deze waarde gelijk aan de correlatie tussen X1 en Y?
- Hoe groot is de determinatiecoëfficiënt voor de regressie van Y op basis van X1?
- Hoe groot is de determinatiecoëfficiënt voor de regressie van Y op basis van X2?
- Deze dataset is een toevalsteekproef. Bereken de betrouwbaarheidsintervallen voor de gemiddelde scores van X1, X2 en Y.
  
- Bereken de partiële correlatiecoëfficiënt tussen X1 en Y onder controle van X2.  
Wat is jouw besluit ?
  
- Hoeveel van de variabiliteit in ‘studieresultaten’(Y) kan verklaard worden op basis van ‘uren kwaliteitstijd met de ouders’ (X1) en ‘aantal keren overwogen om thuis weg te lopen’ (X2) ?
- Welke factor (X1 of X2) heeft relatief het sterkste effect op studieresultaten ?
- Wat is het verwachte studieresultaat voor een adolescent die wekelijks 15u kwaliteitstijd met de ouders doorbrengt en de afgelopen zes maanden geen enkele keer heeft overwogen om thuis weg te lopen ?



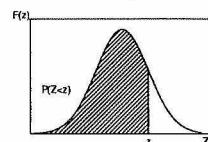
## Referenties

Bij het samenstellen van dit handboek werd dankbaar gebruik gemaakt van volgende bronnen:

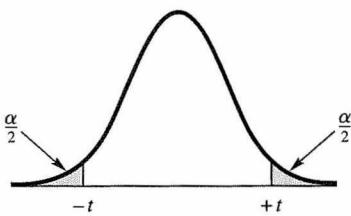
- Achen, C. (1982). *Interpreting and using regression*. Beverly Hills: Sage.
- Brinkman, J. (2009). *Cijfers spreken*. Groningen: Wolters Noordhoff.
- David, F.N. (1962). *Games, gods, and gambling*. New York: Hafner.
- Dudycha, A.L. & Dudycha, L.W. (1972). Behavioral statistics: an historical perspective. In R.E. Kirk (ed.). *Statistical issues: a reader for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.
- Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Fox, J.A., Levin, J. & Forde, D.R. (2009). *Elementary Statistics in Criminal Justice Research*. Boston: Allyn & Bacon.
- Galton, F. (1883). *Inquiries into human faculty and its development*. London: Macmillan.
- Graunt, J. (1662). *Observations in the London bills of mortality*. London: Cambridge.
- Grotenhuis, M. & Van der Wegen, T. (2008). *Statistiek als hulpmiddel. Een overzicht van gangbare toepassingen in de sociale wetenschappen*. Assen: Van Gorcum.
- Kemp, S. & Holmwood, J. (2003). Realism, regularity and social explanation. *Journal of the theory of social behavior*, 33(2), 165-187.
- Loosveldt, G., Maes, C. & Welkenhuysen-Gybels, J. (2008) (8ste druk). *Basisconcepten van de beschrijvende statistiek*. Leuven: Acco.
- McClendon, M. J. (1994). *Multiple regression and causal analysis*. Itasca, IL: FE Peacock Publishers.
- Pauwels, L. (2007). *Buurtinvloeden en jeugddelinquentie: een toets van de sociale desorganisatietheorie*. Den Haag: Boom Juridische Uitgevers.
- Pauwels, L. & Pleysier, P. (2009). *Criminaliteit en onveiligheid meten: de gestandaardiseerde vragenlijst*, Leuven: Acco.
- Pauwels, L. (2012). *Kwantitatieve criminologie. Basishandboek Criminologische Methoden van Criminologisch Onderzoek*.
- Pauwels, L. (2012). *Toegepaste kwantitatieve data-analyse met SPSS voor criminologen*, Antwerpen: Maklu.

- Pearson, K. (1895). Classification of asymmetrical frequency curves in general. Types actually occurring. *Philosophical transactions of the Royal Society of London, Series A*, 186. Cambridge University.
- Petras, H., Masyn, K., Piquero, A. & Weisburd, D. (2010). *Handbook of Quantitative Criminology*. New York: Springer-Verlag.
- Salkind, N. (2008). *Statistics for People Who (Think They) hate Statistics*. Sage, Thousand Oaks.
- Simpson; E.H. (1951). The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, series B*, 13, 238-241.
- Swanborn, P. (2002). *Basisboek sociaal onderzoek*. Amsterdam: Boom.
- Van der Laan. E. (2008). *Basisboek statistiek in business*. Amsterdam: Boom.
- Van Groningen, B., De Boer, C., (2008). *Beschrijvende statistiek: het berekenen en interpreteren van tabellen en statistieken*. Den Haag: Boom.
- Walker, J.T. & Maddan, S. (2013). *Statistics in Criminal Justice: Analysis and Interpretation*. Burlington, MA: Jones & Bartlett.
- Warner, R. M. (2012). *Applied Statistics: From Bivariate Through Multivariate Techniques: From Bivariate Through Multivariate Techniques*. Sage.
- Weisburd, D. & Britt, C. (2007). *Statistics in criminal justice*. Springer.
- Yule, G.U. (1905). The introduction of the words ‘statistics,’ and ‘statistical’ into the English language. *Journal of the Royal Statistical Society*, 68, 391-396.
- Zimmerman, E.A.W. (1787). *A political survey of the present state of Europe*. London: Royal Statistical Society.

Tabel 9.1: Standaard normale verdeling



<i>z</i>	Tweede decimaal van <i>z</i>									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

TABLE C Critical Values of  $t$ 

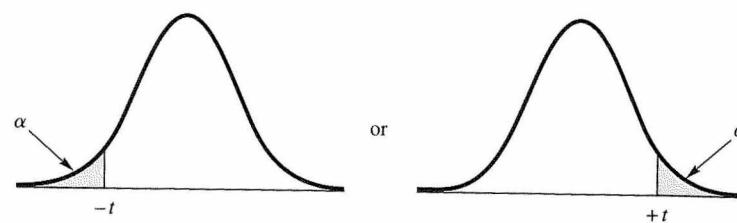
For any given df, the table shows the values of  $t$  corresponding to various levels of probability. Obtained  $t$  is significant at a given level if it is *larger than* the value shown in the table (ignoring the sign).

Level of Significance for Two-Tailed Test ( $\alpha$ )

df	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.291

Note: The bottom row ( $df = \infty$ ) also equals critical values for  $z$ .

TABLE C (continued)

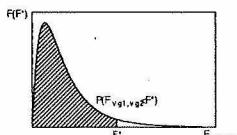


For any given df, the table shows the values of  $t$  corresponding to various levels of probability. Obtained  $t$  is significant at a given level if it is *larger than* the value shown in the table (ignoring the sign).

Level of Significance for One-Tailed Test ( $\alpha$ )

df	.10	.05	.025	.01	.005	.0005
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
$\infty$	1.282	1.645	1.960	2.326	2.576	3.291

Note: The bottom row (df =  $\infty$ ) also equals critical values for  $z$ .



Tabel 9.3: F-verdeling

		Vrijheidsgraden teller										
	P	1	2	3	4	5	6	8	10	20	40	$\infty$
1	.750	5.83	7.50	8.20	8.58	8.82	8.98	9.19	9.32	9.58	9.71	9.85
	.900	39.9	49.5	53.6	55.8	57.2	58.2	59.4	60.2	61.7	62.5	63.3
	.950	161	199	216	225	230	234	239	242	248	251	254
2	.750	2.57	3.00	3.15	3.23	3.28	3.31	3.35	3.38	3.43	3.45	3.48
	.900	8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.39	9.44	9.47	9.49
	.950	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.5	19.5
3	.750	2.02	2.28	2.36	2.39	2.41	2.42	2.44	2.44	2.46	2.47	2.47
	.900	5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.23	5.18	5.16	5.13
	.950	10.1	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.66	8.59	8.53
4	.750	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08
	.900	4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.92	3.84	3.80	3.76
	.950	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.80	5.72	5.63
5	.750	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.87
	.900	4.06	3.78	3.62	3.52	3.45	3.40	3.34	3.30	3.21	3.16	3.11
	.950	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.56	4.46	4.37
6	.750	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.77	1.76	1.75	1.74
	.900	3.78	3.46	3.29	3.18	3.11	3.05	2.98	2.94	2.84	2.78	2.72
	.950	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	3.87	3.77	3.67
7	.750	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.69	1.67	1.66	1.65
	.900	3.59	3.26	3.07	2.96	2.88	2.83	2.75	2.70	2.59	2.54	2.47
	.950	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.44	3.34	3.23
8	.750	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.63	1.61	1.59	1.58
	.900	3.46	3.11	2.92	2.81	2.73	2.67	2.59	2.54	2.42	2.36	2.29
	.950	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.15	3.04	2.93
9	.750	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.59	1.56	1.54	1.53
	.900	3.36	3.01	2.81	2.69	2.61	2.55	2.47	2.42	2.30	2.23	2.16
	.950	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	2.94	2.83	2.71
10	.750	1.49	1.60	1.60	1.59	1.59	1.58	1.56	1.55	1.52	1.51	1.48
	.900	3.29	2.92	2.73	2.61	2.52	2.46	2.38	2.32	2.20	2.13	2.06
	.950	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.77	2.66	2.54
	.990	10.0	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.41	4.17	3.91
	.999	21.0	14.9	12.6	11.3	10.5	9.93	9.20	8.75	7.80	7.30	6.76

Tabel 9.3 voortgezet

	P	1	2	3	4	5	6	8	10	20	40	$\infty$
12	.750	1.46	1.56	1.56	1.55	1.54	1.53	1.51	1.50	1.47	1.45	1.42
	.900	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.19	2.06	1.99	1.90
	.950	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.54	2.43	2.30
	.990	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	3.86	3.62	3.36
	.999	18.6	13.0	10.8	9.63	8.89	8.38	7.71	7.29	6.40	5.93	5.42
14	.750	1.44	1.53	1.53	1.52	1.51	1.50	1.48	1.46	1.43	1.41	1.38
	.900	3.10	2.73	2.52	2.39	2.31	2.24	2.15	2.10	1.96	1.89	1.80
	.950	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.39	2.27	2.13
	.990	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.94	3.51	3.27	3.00
	.999	17.1	11.8	9.73	8.62	7.92	7.44	6.80	6.40	5.56	5.10	4.60
16	.750	1.42	1.51	1.51	1.50	1.48	1.47	1.45	1.44	1.40	1.37	1.34
	.900	3.05	2.67	2.46	2.33	2.24	2.18	2.09	2.03	1.89	1.81	1.72
	.950	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.28	2.15	2.01
	.990	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.26	3.02	2.75
	.999	16.1	11.0	9.01	7.94	7.27	6.80	6.20	5.81	4.99	4.54	4.06
18	.750	1.41	1.50	1.49	1.48	1.46	1.45	1.43	1.42	1.38	1.35	1.32
	.900	3.01	2.62	2.42	2.29	2.20	2.13	2.04	1.98	1.84	1.75	1.66
	.950	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.19	2.06	1.92
	.990	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.51	3.08	2.84	2.57
	.999	15.4	10.4	8.49	7.46	6.81	6.35	5.76	5.39	4.59	4.15	3.67
20	.750	1.40	1.49	1.48	1.47	1.45	1.44	1.42	1.40	1.36	1.33	1.29
	.900	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.94	1.79	1.71	1.61
	.950	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.12	1.99	1.84
	.990	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	2.94	2.69	2.42
	.999	14.8	10.0	8.10	7.10	6.46	6.02	5.44	5.08	4.29	3.86	3.38
30	.750	1.38	1.45	1.44	1.42	1.41	1.39	1.37	1.35	1.30	1.27	1.23
	.900	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.82	1.67	1.57	1.46
	.950	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	1.93	1.79	1.62
	.990	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.55	2.30	2.01
	.999	13.3	8.77	7.05	6.12	5.53	5.12	4.58	4.24	3.49	3.07	2.59
40	.750	1.36	1.44	1.42	1.40	1.39	1.37	1.35	1.33	1.28	1.24	1.19
	.900	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.76	1.61	1.51	1.38
	.950	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	1.84	1.69	1.51
	.990	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.37	2.11	1.80
	.999	12.6	8.25	6.59	5.70	5.13	4.73	4.21	3.87	3.15	2.73	2.23
60	.750	1.35	1.42	1.41	1.38	1.37	1.35	1.32	1.30	1.25	1.21	1.15
	.900	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.71	1.54	1.44	1.29
	.950	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.75	1.59	1.39
	.990	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.20	1.94	1.60
	.999	12.0	7.77	6.17	5.31	4.76	4.37	3.86	3.54	2.83	2.41	1.89
100	.750	1.34	1.41	1.39	1.37	1.35	1.33	1.30	1.28	1.23	1.18	1.11
	.900	2.76	2.36	2.14	2.00	1.91	1.83	1.73	1.66	1.49	1.38	1.21
	.950	3.94	3.09	2.70	2.46	2.31	2.19	2.03	1.93	1.68	1.52	1.28
	.990	6.90	4.82	3.98	3.51	3.21	2.99	2.69	2.50	2.07	1.80	1.43
	.999	11.5	7.41	5.86	5.02	4.48	4.11	3.61	3.30	2.59	2.17	1.62
$\infty$	.750	1.32	1.39	1.37	1.35	1.33	1.31	1.28	1.26	1.19	1.14	1.01
	.900	2.71	2.30	2.08	1.95	1.85	1.77	1.67	1.60	1.42	1.30	1.01
	.950	3.84	3.00	2.61	2.37	2.21	2.10	1.94	1.83	1.57	1.40	1.01
	.990	6.64	4.61	3.78	3.32	3.02	2.80	2.51	2.32	1.88	1.59	1.03
	.999	10.8	6.91	5.43	4.62	4.11	3.75	3.27	2.96	2.27	1.84	1.02

Tabel E Kritieke waarden voor  $F$ -verdeling

		Aantal vrijheidsgraden in de teller								
DFD	p	1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48

Tabel E 707

Tabel E (Vervolg)

Aantal vrijheidsgraden in de teller											
10	12	15	20	25	30	40	50	60	120	1000	
60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30	
241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.19	
968.63	976.71	984.87	993.10	998.08	1001.4	1005.6	1008.1	1009.8	1014.0	1017.7	
6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7	
605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301	
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49	
19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49	
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.49	39.50	
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50	
999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50	
5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13	
8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53	
14.42	14.34	14.25	14.17	14.12	14.08	14.04	14.01	13.99	13.95	13.91	
27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14	
129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53	
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76	
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63	
8.84	8.75	8.66	8.56	8.50	8.46	8.41	8.38	8.36	8.31	8.26	
14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47	
48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09	
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11	
4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37	
6.62	6.52	6.43	6.33	6.27	6.23	6.18	6.14	6.12	6.07	6.02	
10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03	
26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82	
2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72	
4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67	
5.46	5.37	5.27	5.17	5.11	5.07	5.01	4.98	4.96	4.90	4.86	
7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89	
18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77	
2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47	
3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23	
4.76	4.67	4.57	4.47	4.40	4.36	4.31	4.28	4.25	4.20	4.15	
6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66	
14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72	
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30	
3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93	
4.30	4.20	4.10	4.00	3.94	3.89	3.84	3.81	3.78	3.73	3.68	
5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87	
11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36	
2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16	
3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71	
3.96	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34	
5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32	
9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84	
2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06	
2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54	
3.72	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09	
4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92	
8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78	
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98	
2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41	
3.53	3.43	3.33	3.23	3.16	3.12	3.06	3.03	3.00	2.94	2.89	
4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61	
7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02	
2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91	
2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30	
3.37	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73	
4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37	
7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44	

Tabel E (Vervolg)

		Aantal vrijheidsgraden in de teller								
DFD	p	1	2	3	4	5	6	7	8	9
13	0.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	0.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	0.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	0.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	0.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	0.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	0.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	0.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58
15	0.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	0.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	0.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	0.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
16	0.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	0.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	0.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	0.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	0.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	0.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	0.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	0.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	0.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	0.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	0.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	0.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	0.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	0.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	0.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	0.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	0.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	0.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	0.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	0.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	0.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
22	0.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	0.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	0.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	0.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	0.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	0.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	0.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	0.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
24	0.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	0.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	0.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	0.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80

Tabel E 709

Tabel E (Vervolg)

Aantal vrijheidsgraden in de teller										
10	12	15	20	25	30	40	50	60	120	1000
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21
3.25	3.15	3.05	2.95	2.88	2.84	2.78	2.74	2.72	2.66	2.60
4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99
2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
3.15	3.05	2.95	2.84	2.78	2.73	2.67	2.64	2.61	2.55	2.50
3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
6.40	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62
2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
3.06	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33
2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
2.99	2.89	2.79	2.68	2.61	2.57	2.51	2.47	2.45	2.38	2.32
3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76
5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08
2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97
2.92	2.82	2.72	2.62	2.55	2.50	2.44	2.41	2.38	2.32	2.26
3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66
5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87
1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
2.87	2.77	2.67	2.56	2.49	2.44	2.38	2.35	2.32	2.26	2.20
3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
2.38	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
2.82	2.72	2.62	2.51	2.44	2.39	2.33	2.30	2.27	2.20	2.14
3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
2.77	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
2.73	2.64	2.53	2.42	2.36	2.31	2.25	2.21	2.18	2.11	2.05
3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
4.95	4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
2.70	2.60	2.50	2.39	2.32	2.27	2.21	2.17	2.14	2.08	2.01
3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
2.67	2.57	2.47	2.36	2.29	2.24	2.18	2.14	2.11	2.04	1.98
3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
2.64	2.54	2.44	2.33	2.26	2.21	2.15	2.11	2.08	2.01	1.94
3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99

## 710 Tabel E

Tabel E (Vervolg)

		Aantal vrijheidsgraden in de teller								
DFD	p	1	2	3	4	5	6	7	8	9
25	0.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	0.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	0.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
26	0.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	0.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	0.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	0.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
27	0.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	0.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	0.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	0.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57
28	0.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	0.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	0.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	0.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
29	0.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	0.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	0.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	0.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
30	0.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	0.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
40	0.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	0.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	0.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	0.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
50	0.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	0.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	0.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	0.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
60	0.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	0.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	0.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	0.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
100	0.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	0.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
	0.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	0.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	0.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
	0.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	0.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	0.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
	0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
	0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

Tabel E 711

Tabel E (Vervolg)

Aantal vrijheidsgraden in de teller										
10	12	15	20	25	30	40	50	60	120	1000
1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
2.61	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
2.59	2.49	2.39	2.28	2.21	2.16	2.09	2.05	2.03	1.95	1.89
3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
4.48	4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
2.57	2.47	2.36	2.25	2.18	2.13	2.07	2.03	2.00	1.93	1.86
3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78
1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
2.55	2.45	2.34	2.23	2.16	2.11	2.05	2.01	1.98	1.91	1.84
3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
4.35	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
2.53	2.43	2.32	2.21	2.14	2.09	2.03	1.99	1.96	1.89	1.82
3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
4.29	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
2.51	2.41	2.31	2.20	2.12	2.07	2.01	1.97	1.94	1.87	1.80
2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
2.39	2.29	2.18	2.07	1.99	1.94	1.88	1.83	1.80	1.72	1.65
2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
1.73	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
2.32	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
2.70	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
3.67	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
2.27	2.17	2.06	1.94	1.87	1.82	1.74	1.70	1.67	1.58	1.49
2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
2.18	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
1.88	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
2.11	2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
3.12	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
2.06	1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

# Critical Values of $\chi^2$ Distribution

df	$\alpha$					
	0.20	0.10	0.05	0.02	0.01	0.001
1	1.642	2.706	3.841	5.412	6.635	10.827
2	3.219	4.605	5.991	7.824	9.210	13.815
3	4.642	6.251	7.815	9.837	11.341	16.268
4	5.989	7.779	9.488	11.668	13.277	18.465
5	7.289	9.236	11.070	13.388	15.086	20.517
6	8.558	10.645	12.592	15.033	16.812	22.457
7	9.803	12.017	14.067	16.622	18.475	24.322
8	11.030	13.362	15.507	18.168	20.090	26.125
9	12.242	14.684	16.919	19.679	21.666	27.877
10	13.442	15.987	18.307	21.161	23.209	29.588
11	14.631	17.275	19.675	22.618	24.725	31.264
12	15.812	18.549	21.026	24.054	26.217	32.909
13	16.985	19.812	22.362	25.472	27.688	34.528
14	18.151	21.064	23.685	26.873	29.141	36.123
15	19.311	22.307	24.996	28.259	30.578	37.697
16	20.465	23.542	26.296	29.633	32.000	39.252
17	21.615	24.769	27.587	30.995	33.409	40.790
18	22.760	25.989	28.869	32.346	34.805	42.312
19	23.900	27.204	30.144	33.687	36.191	43.820
20	25.038	28.412	31.410	35.020	37.566	45.315
21	26.171	29.615	32.671	36.343	38.932	46.797
22	27.301	30.813	33.924	37.659	40.289	48.268
23	28.429	32.007	35.172	38.968	41.638	49.728
24	29.553	33.196	36.415	40.270	42.980	51.179
25	30.675	34.382	37.652	41.566	44.314	52.620
26	31.795	35.563	38.885	42.856	45.642	54.052
27	32.912	36.741	40.113	44.140	46.963	55.476
28	34.027	37.916	41.337	45.419	48.278	56.893
29	35.139	39.087	42.557	46.693	49.588	58.302
30	36.250	40.256	43.773	47.962	50.892	59.703

Source: From Table IV of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (London: Longman Group Ltd., 1974). (Previously published by Oliver & Boyd, Edinburgh.) Reprinted by permission of Pearson Education Ltd.