

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 II

OEFENBOEK

Prof. dr. Lieven Pauwels
dr. Ann De Buck



Copyright © 2025: Lieven Pauwels

Auteurs: Lieven Pauwels & Ann De Buck

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

Beste student,

Voor jou ligt *Deel II van de Basiscursus Statistiek in de Criminologie: Oefenboek* voor het academiejaar 2025-2026. Dit oefenboek ondersteunt je bij het begrijpen én toepassen van de basisprincipes van de univariate, bivariate en inferentiële statistiek. Elk hoofdstuk bevat een overzicht van de belangrijkste theoretische en methodologische concepten, uitgewerkte praktijkvoorbeelden en oefeningen die je tijdens de werkcolleges zelf zal maken.

Door deze cursus ontdek je hoe statistiek wordt ingezet in de criminologische praktijk om onderzoeks vragen te beantwoorden. Sommige statistische maten komen zo vaak voor dat je ze zelf vlot handmatig moet kunnen berekenen, zoals het gemiddelde, de variantie, de standaardafwijking, de covariantie, de correlatie en regressieanalyses. Ook technieken zoals de chi-kwadraattoets, betrouwbaarheidsintervallen en variantie-analyse dien je zelfstandig te kunnen berekenen en interpreteren. Hiervoor dien je een niet-grafische rekenmachine te gebruiken. Andere statistische methoden moet je vooral leren begrijpen en correct interpreteren. Een grondige kennis van kansverdelingen – in het bijzonder de normale verdeling – is hierbij onmisbaar. Even belangrijk is dat je leert statistische verbanden op waarde te schatten en gefundeerde beslissingen te nemen bij het aannemen of verwijderen van onderzoeksresultaten. Daarbij hoort ook het formuleren en toetsen van hypothesen, en inzicht in de implicaties van het verwijderen van de nulhypothese op basis van echte criminologische data.

Dit oefenboek is echter meer dan een verzameling berekeningen. Omdat het werk veld veel belang hecht aan praktische toepassingen, bevat het ook een inleiding tot SPSS, een veelgebruikt softwarepakket voor data-analyse. Met SPSS genereer je tabellen, grafieken en coëfficiënten, maar de interpretatie van de output blijft mensenwerk. Daarom vind je in de appendix praktijkvoorbeelden waarin de procedures worden toegelicht en vooral de interpretatie van de resultaten centraal staat. De klemtoon ligt daarbij steeds op de inhoudelijke betekenis voor criminologisch onderzoek. Dit oefenboek bereidt je voor op de vervolg cursus *Kwantitatieve Toegepaste Data-analyse en Rapportage* in het tweede bachelorjaar Criminologische Wetenschappen (vanaf academiejaar 2026-2027).

Hoewel we de wiskunde tot een minimum beperken, blijft het uitvoeren van enkele berekeningen noodzakelijk om de basisprincipes goed te begrijpen. Enige elementaire wiskundige kennis is dus vereist. Op de UFORA-cursussite M4S: Maths for Stats vind je kenniscips en zelftests om je basisvaardigheden in wiskunde te versterken en te toetsen.

Tot slot: wil je slagen voor dit opleidingsonderdeel, dan is regelmatig oefenen (lees: elke dag!) onontbeerlijk. Statistiek leer je door afwisselend theorie te bestuderen en oefeningen te maken, waarna je teruggrijpt naar de theorie om je begrip verder te verdiepen. Herhaal dit proces totdat je zowel de theorie als de toepassingen volledig beheerst. Dit handboek helpt je om opdrachten voor te bereiden die in kleinere groepen tijdens de werkcolleges worden verbeterd.

Wij wensen je veel succes en vooral veel leerrendement toe!

Augustus, 2025

Prof. dr. Lieven Pauwels & dr. Ann De Buck

Verantwoordelijk lesgever

Prof. dr. Lieven J.R. Pauwels

Vakgroep Criminologie, Strafrecht en Sociaal Recht

Universiteitstraat 4-6

9000 Gent

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

E-mail: Lieven.Pauwels@ugent.be

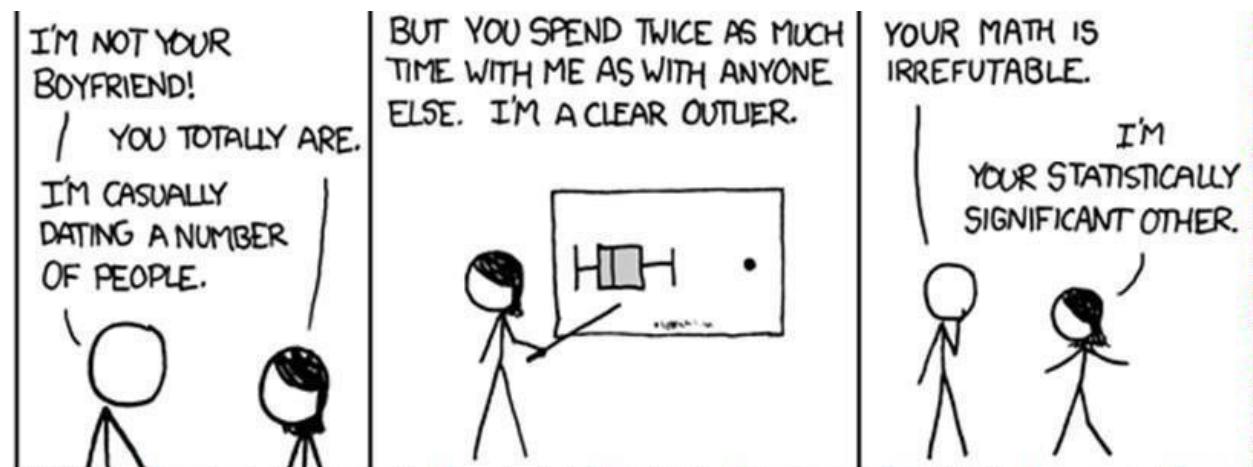
Medelesgever

dr. Ann De Buck

Vakgroep Criminologie, Strafrecht en Sociaal Recht (IRCP)

Tel: ++32 (0)9 264 84 75

E-mail: ann.debuck@ugent.be (bureel: derde verdieping)



INHOUDSOPGAVE

VOORWOORD

HOOFDSTUK 1 - BASISCONCEPTEN	1-16
HOOFDSTUK 2 – UNIVARIATE BESCHRIJVENDE STATISTIEK	17-54
1. Doelstellingen	
2. Frequenties	
2.1. Te onthouden kernbegrippen	
3. Maten van centraliteit	
3.1. Te onthouden kernbegrippen	
3.2. Statistische symbolen en formules	
4. Maten van spreiding	
4.1. Te onthouden kernbegrippen	
4.2. Statistische symbolen en formules	
5. Samenvattende tabel	
6. Maten van vorm	
6.1. Te onthouden kernbegrippen	
6.2. Statistische symbolen en formules	
7. Oefeningen	
8. Extra oefeningen	
HOOFDSTUK 3 – TOEPASSINGEN OP KANSREKENEN	55-58
1. Inleiding	
2. Uitspraken	
3. Oefeningen	
HOOFDSTUK 4 – STANDAARDNORMALE VERDELING	59-64
1. Doelstellingen	
2. Te onthouden kernbegrippen	
3. Statistische symbolen en formules	
4. Oefeningen	
HOOFDSTUK 5 – BIVARIATE STATISTIEK	65-110
1. Doelstellingen	
2. Bivariate associatiematen tussen categorische variabelen	
2.1. Nominale variabelen	
2.1.1. Te onthouden kernbegrippen	
2.1.2. Statistische symbolen en formules	
2.1.3. Associatiematen op nominaal niveau: SPSS-output en interpretatie	
2.2. Ordinale variabelen	
2.2.1. Te onthouden kernbegrippen	
2.2.2. Associatiematen op ordinaal niveau: SPSS-output en interpretatie	

3. Bivariate associatiematen tussen metrische variabelen
 - 3.1. Symmetrische associatiematen: correlatie analyse
 - 3.1.1. Te onthouden kernbegrippen
 - 3.1.2. Statistische symbolen en formules
 - 3.1.3. Symmetrische associatiematen op metrisch niveau:
SPSS-output en interpretatie
 - 3.1.4. Asymmetrische associatiematen op metrisch niveau:
SPSS-output en interpretatie
 4. Samenvattende tabel
 5. Oefeningen

HOOFDSTUK 6 – INFERENTIELE STATISTIEK	111-142
1. Doelstellingen	
2. Te onthouden kernbegrippen	
3. Statistische symbolen en formules	
4. Centrale limietstelling: didactisch voorbeeld	
5. Toepassen inferentiële statistiek in de praktijk	
6. Oefeningen	
HOOFDSTUK 7 - VARIANTIE ANALYSE	143-154
1. Doelstellingen	
2. Te onthouden kernbegrippen	
3. Statistische symbolen en formules	
4. Oefeningen	
HOOFDSTUK 8 - PARTIELE CORRELATIE	155-162
1. Doelstellingen	
2. Te onthouden kernbegrippen	
3. Statistische symbolen en formules	
4. Oefeningen	
HOOFDSTUK 9 - MULTIPLE REGRESSIE ANALYSE	163-176
1. Doelstellingen	
2. Te onthouden kernbegrippen	
3. Statistische symbolen en formules	
4. Oefeningen	
HOOFDSTUK 10 - PADMODEL_INLEIDING	177-184
1. Doelstellingen	
2. Te onthouden kernbegrippen	
3. Oefeningen	
HOOFDSTUK 11 - SYNTHESE OEFENING TER ILLUSTRATIE VAN EXAMEN	185-188
HOOFDSTUK 12 - HERHALING	189-202

INHOUDSOPGAVE

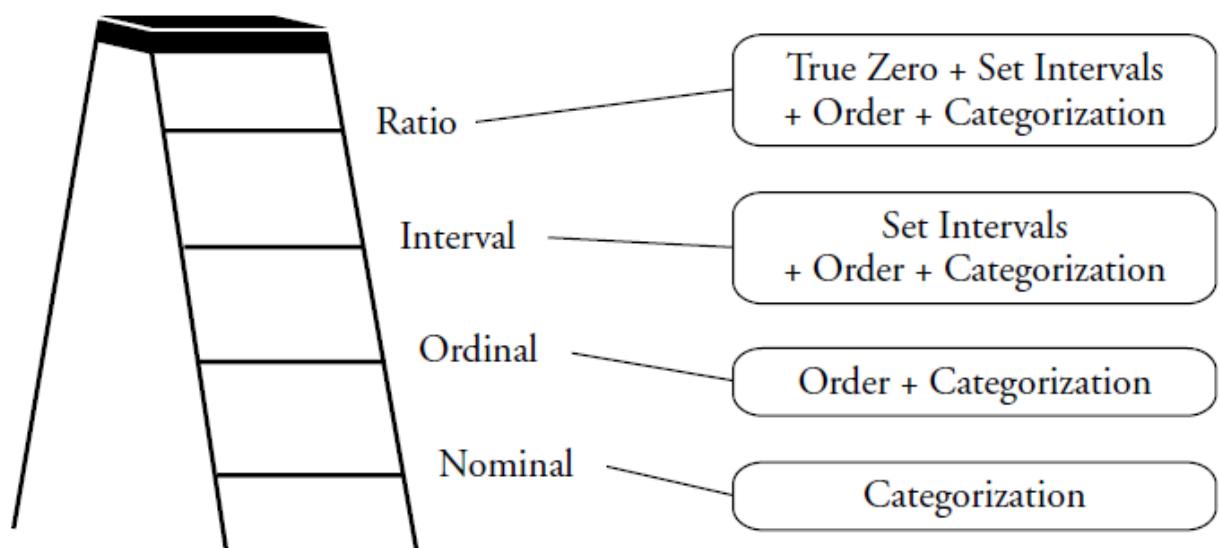
BIBLIOGRAFIE	203
OPLOSSING SYNTHESE OEFENING	204-214
APPENDIX_INLEIDING TOT SPSS	215-257

INHOUDSOPGAVE

HOOFDSTUK 1

BASISCONCEPTEN

Ladder of Measurement



(bron: Weisburd & Britt, 2007: 16)

HOOFDSTUK I

BASISCONCEPTEN

1. Afrondingsregels: rond af tot op twee decimalen

- | | |
|----------|----------|
| * 6.978 | * 0.915 |
| * 10.657 | * 3.879 |
| * 87,001 | * 0,5555 |
| * 55,253 | * 0,6565 |
| * 7,523 | |
| * 20.954 | |

2. Meetniveaus en centrale concepten.

Hieronder volgen een aantal vragen uit een vragenlijst die gebruikt werden in een onderzoek naar de rol van specifieke variabelen uit verschillende criminologische theorieën ter verklaring van 'intenties om te stelen'.

Identificeer het meetniveau.

- a. Ik kan verleidingen goed weerstaan.
 - Helemaal akkoord
 - Akkoord
 - Nog akkoord noch niet akkoord
 - Niet akkoord
 - Helemaal niet akkoord
- b. Hoe waarschijnlijk is het dat je betrapt wordt?
 - Heel waarschijnlijk
 - Waarschijnlijk
 - Neutraal
 - Onwaarschijnlijk
 - Heel onwaarschijnlijk
- c. Heb jij Ooit al eens geld gestolen (om het even welk bedrag)?
 - Nee
 - Ja

- d. Hoe vaak heb je dat in de laatste 12 maanden gedaan?
- 1 keer
 - 2 keer
 - 3 keer
 - 4-6 keer
 - Meer dan 6 keer
- e. Hoe verkeerd is het volgens jou dat iemand van jouw leeftijd iemands portefeuille steelt?
- Helemaal verkeerd
 - Verkeerd
 - Nog verkeerd noch niet verkeerd
 - Niet verkeerd
 - Helemaal niet verkeerd
- f. Op een schaal van 1 tot 10 hoe eerlijk denk je dat jij bent, vergeleken met andere personen die deze vragenlijst hebben ingevuld?
- (vul hier een score in)
- g. Wat is jouw geslacht bij geboorte?
- Jongen
 - Meisje
 - Zeg ik liever niet
- h. In welk jaar ben je geboren?
- (geboortejaar)
- i. Tot welke geloofsgemeenschap reken je jezelf?
- Ik behoor niet tot een geloofsgemeenschap
 - Katholiek
 - Protestant
 - Soennitische Islam
 - Sjiitische Islam
 - Joods
 - Een andere geloofsovertuiging

j. Hoe vaak bid jij?

- Elke dag
- Meer dan 1 keer per week
- 1 keer per week
- 1 keer per maand
- Alleen op speciale religieuze dagen
- Nooit

4. Hieronder vind je een aantal variabelen die in criminologisch onderzoek vaak voorkomen. Maak een voorstel voor nominale – ordinale en metrische variabelen.

Maak van onderstaande variabelen een variabele van het NOMINALE meetniveau	
Variabele	Antwoordcategorieën
Biologische sexe	
Burgerlijke staat	
Voorwaardelijke hechtenis	
Type aanklacht	
Type straf	
Maak van onderstaande variabelen een variabele van het ORDINALE meetniveau	
Variabele	Antwoordcategorieën
Opleidingsniveau	
Ernst van verwondingen bij een aanval	
Attitude in een opinievraag	
Type straf	
<i>"Bent u persoonlijk bezorgd over de criminaliteit in Gen?"</i>	

Maak van onderstaande variabelen een variabele van het METRISCHE meetniveau	
Variabele	Antwoordcategorieën
Leeftijd	
Opleiding	
Inkomen	
Zelf-gerapporteerde diefstal	
Diefstallen in de stad Gent	
Sociale contacten in de buurt	

5. Geef voor elk van de volgende voorbeelden van strafrechtelijk onderzoek aan of de gebruikte meetschaal nominaal, ordinaal of ten minste interval is (d.w.z. interval of ratio). Leg je keuze uit.
- a. In een huis-aan-huis onderzoek wordt aan de bewoners van een buurt gevraagd hoe vaak zij (of iemand in hun huishouden) het afgelopen jaar slachtoffer zijn geweest van een misdrijf.
 - b. Justitie-assistenten beoordelen het gedrag van gedetineerden op een schaal met waarden variërend van 1 tot 10; een score van 1 staat voor voorbeeldig gedrag.
 - c. Aan honderd studenten wordt gevraagd of ze ooit gearresteerd zijn.
 - d. Een onderzoeker controleert gevangenisgegevens om het land van herkomst te bepalen van gevangenen in een bepaald cellenblok.
 - e. In een telefonische enquête wordt aan het publiek gevraagd welke van de volgende antwoordopties het beste overeenkomt met hoe ze denken over de dienstverlening van de locale politie: heel erg ontevreden – ontevreden – neutraal – tevreden – heel tevreden.
 - f. Een criminoloog meet de diameter (in centimeters) van de schedels van gevangenen die in de gevangenis zijn gestorven, in een poging om een biologische theorie te ontwikkelen over de oorzaken van criminaliteit.
 - g. Aan secretaresses van een topadvocatenkantoor wordt de volgende vraag gesteld: "*Bent u het afgelopen jaar het slachtoffer geweest van seksuele intimidatie en indien ja, hoeveel keer?*" Antwoordopties waren : nooit – één keer – twee of drie keer – meer dan drie keer – weiger te antwoorden.

6. Een onderzoeksteam werd ingehuurd om een rehabilitatieprogramma te testen dat als doel heeft de kans op recidive van ex-gedetineerden na hun vrijlating te verminderen. 400 gedetineerden kwamen in aanmerking voor het onderzoek. De helft werd op basis van toeval aangeduid voor deelname aan het programma, de andere helft bleef in de gevangenis zonder deelname aan het rehabilitatieprogramma. Na hun vrijlating werden alle ex-gedetineerden gedurende 2 jaar gevolgd om na te gaan in welke mate zij opnieuw delicten pleegden. Het onderzoeksteam vergeleek daarbij het aantal arrestaties van de groep die niet en de groep die wel aan een rehabilitatieprogramma had deelgenomen.
- a. Wat is hier de onafhankelijke variabele ?
 - b. Wat is het meetniveau van de onafhankelijke variabele ?
 - c. Wat is hier de afhankelijke variabele ?
 - d. Wat is het meetniveau van de afhankelijke variabele ?

7. In een onderzoek onder 552 eerstejaarsstudenten werd gevraagd naar een favoriete keuze om online onderzoek te doen.

Mogelijke keuzes waren: Google GoogleScholar Wikipedia Andere.

De namen van de studenten werden niet genoteerd maar de studenten kregen een volgnummer in het databestand van 1 tot 552.

De onderzoekers registreerden ook leeftijd – geslacht – en studierichting van elke student.

- Wat/wie zijn de onderzoekseenheden?
- Wat zijn de onderzochte variabelen?
- Wat is het meetniveau?

8. Ontleed de volgende uitspraken en antwoord telkens op de volgende vragen:

- a. Wat is de bestudeerde populatie ?
- b. Wat is 1 statistische eenheid ?
- c. Wat zijn de bestudeerde variabelen ?
- d. Op welk meetniveau zijn de variabelen gemeten ?

"Uit de resultaten van de studentenbarometer – een vragenlijst verstrekt aan alle studenten van de UGent – is gebleken dat studenten aan de faculteit Recht en Criminologie gemiddeld 2.5 glazen bier per week meer drinken dan studenten aan de faculteit Politieke en Sociale Wetenschappen"

"Steeds meer jongeren grijpen naar wapens om zich veilig te voelen op school. Dat blijkt uit een enquête gehouden bij 239 leerlingen van verschillende Antwerpse scholen. 17 procent van de ondervraagde jongens en 7 procent van de meisjes tussen 14 en 17 jaar komt er openlijk voor uit al eens een wapen mee naar school te nemen. Jongens hebben meestal een mes op zak, meisjes een busje pepperspray"

"Een groepje criminologiestudenten bestudeert verkeersagressie. Ze tellen het totaal aantal PV's opgemaakt voor verkeersagressie voor elk Vlaams arrondissement."

"In de Verenigde Staten kunnen herfststormen nog wel eens lelijk huis houden. Maar niet elke staat heeft daar even veel last van. Om na te gaan welke staten het meest en het minst geconfronteerd worden met deze stormen, werd voor elke herfststorm in het jaar 2014 nagegaan waar die plaatsvonden."

9. Hoe zou jij universiteiten beoordelen?

Beschrijf vijf variabelen die je zou willen meten voor elke universiteit als je zou kiezen waar je gaat studeren.

Bepaal het meetniveau – Wat zijn antwoordopties?

VARIABELE	MEETNIVEAU	ANTWOORDOPTIES

10. Hieronder vind je een deel uit een databestand.

De kolommen vertegenwoordigen variabelen/kenmerken – de rijen vertegenwoordigen statistische eenheden ($n=7$).

Wat is het meetniveau van elke variabele?

- Type restaurant
- Naam restaurant
- Menu
- Prijzen_normaal
- Prijzen_korting

	TYPE_RESTAURANT	NAAM_RESTAURANT	MENU	PRIJZEN_NORMAAL	PRIJZEN_KORTING
1	Italiaans	Domo	Pizza	20	10
2	Italiaans	Mama Rita	Pizza	20	12
3	Grill	Smokey McQueen	Barbecue	30	17
4	Grill	Smokey Grill	Ribbetjes	20	11
5	Mexicaans	Dos Amigos	Tacos	16	8
6	Mexicaans	Holy Guacamole	Steak	13	8
7	Vis	Sea Grill	Garnalen	20	11

11. Op welk meetniveau werden volgende variabelen gemeten ? Geef bij metrische/kwantitatieve variabelen aan of het om een discrete of continue variabele gaat.

Studentenbarometer 2008-2009

33. /50. Alcohol II

1. In welke mate ga je akkoord dat drinken:

	Helemaal akkoord	Akkoord	Neutraal	Niet akkoord	Helemaal niet akkoord
1. Je ontspant na de schooluren	<input type="radio"/>				
2. De schoolstress doet dalen	<input type="radio"/>				
3. Helpt schoolproblemen te vergeten	<input type="radio"/>				
4. De werkdruk doet afnemen	<input type="radio"/>				
5. Helpt relaxen en ontspannen tijdens het weekend	<input type="radio"/>				
6. Je helpt op te beuren als je depressief of verdrietig bent	<input type="radio"/>				

2. Gedurende de voorbije 7 dagen, hoeveel glazen bier, wijn en sterke drank heb je gedronken?

1. Bier

2. Wijn

3. Sterke drank

Studentenbarometer 2008-2009

17. /50. Geboortemaand

1. In welke maand ben je geboren?

- Januari
- Februari
- Maart
- April
- Mei
- Juni
- Juli
- Augustus
- September
- Oktober
- November
- December

Studentenbarometer 2008-2009

2. /50. Demografische achtergrond

1. Geslacht

- mannelijk
- vrouwelijk

2. Ik ben geboren in het jaar...

3. Hoeveel broers en zussen heb je (inclusief adoptie- en stiefkinderen)?

1. Aantal broers

2. Aantal zussen

4. Heb je een functiebeperking?

- Ja, ik heb een functiebeperking
- Nee, ik heb geen functiebeperking

Studentenbarometer 2008-2009

16. /50. Woonsituatie

1. Waar verblijf je tijdens de week?

- Op een studentenhome van de universiteit
- In een kamer in een private woning
- In een studio of appartement
- Ik huur een huis of appartement samen met anderen
- Thuis

2. Wat is het postnummer van de plaats waar je gedomicileerd bent?

Studentenbarometer 2008-2009

4. /50. Middelbare studies

1. In welke onderwijsvorm zat je in het zesde middelbaar?

- algemeen secundair onderwijs (ASO)
- kunstsecundair onderwijs (KSO)
- technisch secundair onderwijs (TSO)
- beroepssecundair onderwijs (BSO)

2. Welk totaalpercentage behaalde je op het einde van het zesde middelbaar?

12. Zijn volgende uitspraken juist of fout ?

Uitspraak	JUIST	FOUT
Hoe lager het meetniveau, hoe gedetailleerder de informatie.		
Als een variabele voor elke eenheid dezelfde waarde aanneemt, dan is deze een constante.		
Een discrete meetschaal is een meetschaal waarbij alle tussenliggende waarden tussen twee waarden kunnen voorkomen.		
Als een variabele gemeten wordt op intervalniveau, kan je die ook herleiden naar een ordinaal niveau ten behoeve van een statistische analyse.		
Het intervalniveau heeft een vast nulpunt.		
De multivariate statistiek houdt zich bezig met de relatie tussen twee variabelen.		
De inferentiële statistiek houdt zich bezig met de beschrijving van fenomenen.		
Een verklarende onderzoeksvraag beantwoorden houdt minstens bivariate statistiek in.		
Een beschrijvende onderzoeksvraag beantwoorden kan gebeuren aan de hand van univariate en bivariate statistiek.		
Wanneer we een criminologische theorie willen toetsen, moeten we gebruik maken van de inferentiële statistiek.		
Nominaal meetniveau is een meetniveau dat informatie categoriseert en een orde van grootte toekent zonder gebruik te maken van een standaardschaal met gelijke intervallen.		

13. Beschrijf het verschil tussen:

- a. Beschrijvende en verklarende statistiek
 - b. Een populatie en een variabele
 - c. Een populatie en een steekproef

14. Formuleren van onderzoeks vragen

- a. Formuleer een criminologisch relevante univariate beschrijvende onderzoeks vrag.

Wat is de variabele ?

Wat is het meetniveau ?

Welke waarden kan deze variabelen aannemen ?

- b. Formuleer een criminologisch relevant bivariate beschrijvende onderzoeks vrag.

Wat zijn de variabelen ?

Wat zijn de meetniveaus ?

Welke waarden kunnen deze variabelen aannemen ?

- c. Formuleer een criminologisch relevante bivariate verklarende onderzoeks vrag.

Wat zijn de variabelen ?

Wat zijn de meetniveaus ?

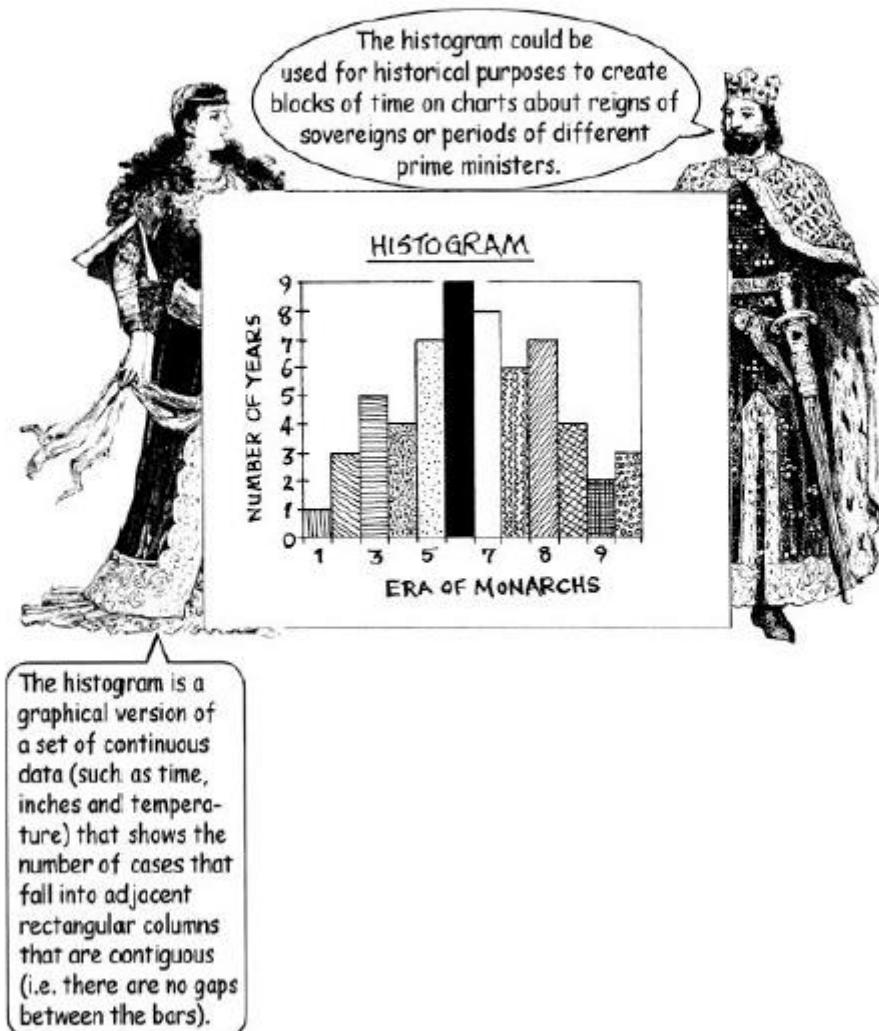
Welke waarden kunnen deze variabelen aannemen ?

HOOFDSTUK 2

UNIVARIATE BESCHRIJVENDE STATISTIEK

The Histogram

Pearson introduced the histogram on 18 November 1891. It is a term that he coined to designate a “time-diagram” in his lecture on “Maps and Chartograms”.



(bron: Magnello & Van Loon, 2009)

HOOFDSTUK II

UNIVARIATE BESCHRIJVENDE STATISTIEK

1. DOELSTELLINGEN

Op het einde van dit hoofdstuk zijn studenten in staat om onderzoeksgegevens samen te vatten, te analyseren en te interpreteren. Daarbij dienen zij rekening te houden met het meetniveau en de behoeften van de analyse. Studenten kunnen frequentietabellen en grafieken correct lezen en interpreteren.

2. FREQUENTIES

2.1. Te onthouden kernbegrippen

Absolute frequentie	Het aantal keer dat een bepaalde waarde of categorie van een variabele voorkomt.
Cumulatieve frequentie of percentage	Kan pas berekend worden als je categorieën logisch kan ordenen van laag naar hoog. Bij elke frequentie of percentage tellen we de volgende op. De cumulatieve frequentie van de laatste waarde is steeds gelijk aan het totale aantal waarnemingen (idem voor het cumulatief percentage bij de laatste waarde). Hierdoor is het mogelijk te zien hoeveel waarnemingen kleiner dan of gelijk zijn aan een bepaalde waarde.
Frequentie	Het aantal keer dat een score of waarde voorkomt
Frequentieverdeling	Rangschikking van het aantal waarnemingen van laag naar hoog met vermelding van het aantal keer dat elke categorie van de variabele voorkomt.
Geldig percentage	Percentage berekend op diegenen die een geldig antwoord gaven op de vraag
Histogram	Grafische voorstelling van een frequentieverdeling van metrische variabelen. Lijkt op een staafdiagram maar de staafjes worden aan elkaar getekend omdat de waarden van de variabele elkaar opvolgen. De categorieën liggen op een continuüm.
Percentage	De verhouding tussen het absolute aantal keer dat een waarde voorkomt gedeeld door de totale steekproefgrootte (n) maal 100
Proportie	Absolute frequentie delen door het totaal aantal waarnemingen (n), getal dat gaat van 0 tot 1. De som van alle proporties is gelijk aan 1.
Relatieve frequentie	Geeft weer hoe vaak een waarde voorkomt ten opzichte van het totaal aantal waarnemingen. Kan worden uitgedrukt in

	percentage op 100 of in proporties (of fracties) op 1 (ten opzichte van de geldige steekproefomvang)
Staafdiagram	De gegevens worden voorgesteld aan de hand van twee assen. Op de horizontale as (X-as) worden de verschillende categorieën van de variabele weergegeven. De verticale as (Y-as) geeft de aantallen weer onder de vorm van absolute frequenties of percentages. Bij elke categorie op de X-as tekent men een staafje en deze worden los van elkaar geplaatst. Geschikt voor grafische voorstelling van variabelen gemeten op het nominale en ordinale niveau.
Taartdiagram of cirkeldiagram	Een cirkelschijf verdeeld in sectoren op basis van de verschillende frequenties of percentages. Er wordt gebruik van gemaakt wanneer we met een beperkt aantal categorieën werken. Geschikt voor variabelen gemeten op het nominale en ordinale niveau.



! Vooraleer een univariate beschrijvende analyse uit te voeren, dien je jezelf volgende vragen te stellen:

1. *Wat is het gemeten kenmerk/wat is de variabele*
2. *Uit welke categorieën bestaat de variabele*
3. *Wat is het meetniveau van de bestudeerde variabele*
4. *welke beschrijvende onderzoeksvervraag wil je beantwoorden: een vraag naar centraliteit, spreiding of vorm van een kenmerk*



! Leer goed het onderscheid tussen exacte frequenties, percentage en geldige percentages. We bespreken enkel de geldige percentages. Je dient te vermelden op basis van hoeveel geldige antwoorden (= aantal respondenten dat de vraag geldig beantwoordde) de onderzoeksvervraag werd beantwoord. Vermeld ook steeds het totale aantal respondenten. Op die manier heeft de lezer inzicht in het aantal missings of het aantal respondenten dat de vraag niet of niet geldig beantwoordde.

3. MAREN VAN CENTRALITEIT

3.1. Te onthouden kernbegrippen

Deviatiescore	Afwijking ten opzichte van het gemiddelde. De som van de deviatiescores is steeds 0.
Kwantielen	Een maat die men gebruikt wanneer men een geordende rij elementen in om het even welk aantal groepen met een gelijk aantal elementen verdeelt. Ze kunnen berekend worden vanaf het ordinale niveau.
Kwartielen	De drie waarden (Q1, Q2 en Q3) die een geordende reeks uitkomsten in 4 gelijke stukken verdelen, die ieder een kwart van de uitkomsten bevatten.
Maten van centraliteit of centrummaten	Geven weer rond welke waarden de verdeling van een variabele gepositioneerd is. Maten van centraliteit geven weer welke de meest centrale waarde is van een verdeling. Welke maat van centraliteit je kan gebruiken, hangt af van het meetniveau van de variabele. Geven een antwoord op beschrijvende onderzoeks vragen die erop gericht zijn centrale tendensen te ontdekken.
Mediaan	Het midden van een statistische verdeling. Het is een centrummaat die het punt in de frequentieverdeling aangeeft waaronder 50% van de gevallen en waarboven de andere 50% van de gevallen liggen. Kortom, de frequentieverdeling wordt in twee gelijke stukken gedeeld. Om de mediaan te bepalen moeten de categorieën in oplopende volgorde gerangschikt zijn. Ze kan berekend worden vanaf het ordinale niveau.
modus	De modus is een centrummaat: de categorie van de variabele met de hoogste frequentie. Kan gehanteerd worden voor alle meetniveaus.
Rekenkundig gemiddelde	Een centrummaat die wordt verkregen door alle voorkomende waarden bij elkaar op te tellen en vervolgens het totaal te delen door het aantal respondenten. Ze wordt gebruikt bij variabelen gemeten op het metrische niveau.

4. MATEN VAN SPREIDING

4.1. Te onthouden kernbegrippen

Gemiddelde absolute afwijking	Een metrische spreidingsmaat. Het is de som van de absolute waarden van de afwijkingen van elke waarde ten aanzien van het rekenkundig gemiddelde, gedeeld door het aantal waarnemingen.
Index van diversiteit (ID)	Nominale spreidingsmaat, gebaseerd op de relatieve frequenties van de categorieën. Er wordt rekening gehouden met de proportie van waarnemingen binnen iedere categorie. De waarde van deze parameter geeft een idee van de mate van concentratie van de waarnemingen over de categorieën van de variabele.
Interkwartielafstand (K3-K1)	Ordinalle spreidingsmaat. De interkwartielafstand is het verschil tussen de derde en eerste kwartiel en bevat de helft van het totale aantal waarnemingen.
Spreidingsmaat	Geeft aan of waarden in een verdeling dicht bij elkaar liggen of juist ver uit elkaar
Standaardafwijking (s)	De vierkantswortel van de steekproefvariantie
Variatie (SS)	Metrische spreidingsmaat. Het is de som van de gekwadrateerde afwijkingen van elke waarde tegenover het gemiddelde.
Variantie (s^2)	De som van de gekwadrateerde afwijkingen van elke observatie tegenover het gemiddelde, gedeeld door $n-1$. De variantie is dus gelijk aan de variatie delen door $n-1$.
Variatiebreedte	Ordinalle spreidingsmaat. De variatiebreedte is het verschil tussen de grootste en de kleinste waargenomen waarde.
Variatiecoëfficiënt (v)	Gestandaardiseerde spreidingsmaat. Wordt berekend door de standaardafwijking te delen door het rekenkundig gemiddelde. Niet afhankelijk van de meeteenheid en laat bijgevolg toe de spreidingen van verdelingen die worden uitgedrukt in verschillende meeteenheden met elkaar te vergelijken.
Variatieratio (VR)	Een nominale spreidingsmaat. De variatieratio is de proportie waarnemingen die niet tot de modale categorie behoren. Ze neemt de waarde nul aan indien alle waarnemingen tot de modale categorie behoren. Er is echter geen vaste bovengrens.

4.2. Statistische symbolen en formules

Mediaan	$(n + 1)/2$
Rekenkundig gemiddelde	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ <p style="text-align: center; margin-left: 100px;"> het rekenkundig gemiddelde de som van alle waarnemingen </p> <p style="text-align: right; margin-right: 100px;"> het aantal waarnemingen </p>
Som van de deviatiescores = 0	$\sum_{i=1}^n (x_i - \bar{x}) = 0$
Standaardafwijking	$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$ of $s = \sqrt{s^2}$
Variatie	$SS = \sum (X - \bar{X})^2$ of $SS = \sum_{i=1}^N (x_i - \bar{x})^2$
variantie	$s^2 = \frac{\sum (X - \bar{X})^2}{n-1}$
variatiecoëfficiënt	$v = \frac{s}{\bar{X}}$

5. SAMENVATTENDE TABEL : maten van centraliteit en spreiding

	NOMINAAL	ORDINAAL	METRISCH
MATEN VAN CENTRALITEIT	Modus	Modus Mediaan kwartielën	Modus Mediaan Kwartielën Rekenkundig gemiddelde
MATEN VAN SPREIDING	Variatieratio (VR) Index van diversiteit	Variatiebreedte Interkwartielafstand (K3-K1)	Gemiddelde absolute afwijking Variatie (SS) Variantie (s^2) Steekproefstandaardafwijking (s) Variatiecoëfficiënt (v)

6. MATEN VAN VORM

6.1. Te onthouden kernbegrippen

Boxplot	Visuele weergave van centraliteit, spreiding en vorm van de verdeling (voor kenmerken vanaf ordinaal meetniveau).
Empirische coëfficiënt van Pearson	Parameter van vorm. Wordt berekend als volgt: 1) verschil tussen gemiddelde en mediaan en 2) deze waarde delen door de standaardafwijking. Positieve coëfficiënt duidt op een positief asymmetrische verdeling, Negatieve coëfficiënt duidt op een negatief asymmetrische verdeling.
Kurtosis of afplatting	Mate van afplatting van gegevens rondom het rekenkundige gemiddelde.
Negatief asymmetrische verdeling	Verdeling met een langere staart naar links: het rekenkundig gemiddelde is kleiner dan de mediaan. Lage waarden komen minder vaak voor dan hoge waarden.
Positief asymmetrische verdeling	Verdeling met een langere staart naar rechts: het rekenkundig gemiddelde is groter dan de mediaan. Hoge waarden komen minder vaak voor dan lage waarden
Symmetrische verdeling	Verdeling van gegevens ten opzichte van het rekenkundig gemiddelde. Een verdeling is symmetrisch als het rekenkundig gemiddelde en de mediaan aan elkaar gelijk zijn. Een symmetrische verdeling volgt een Gauss-curve, linker- en rechterhelft zijn elkaar spiegelbeeld.

6.2. Statistische symbolen en formules

Empirische coëfficiënt van Pearson	$S = \frac{\bar{X} - \tilde{x}}{s}$
------------------------------------	-------------------------------------

7. OEFENINGEN

1. Wanneer we scores rangschikken volgens hun waarde en frequentie, construeren we een zogenaamde *frequentieverdeling*. Hieronder vind je de gegevens over eerdere arrestaties van 100 bekende overtreders:

14	0	34	8	7	22	12	12	2	8
6	1	8	1	18	8	1	10	10	2
12	26	8	7	9	9	3	2	7	16
8	65	8	2	4	2	4	0	7	2
1	2	11	2	1	1	5	7	4	10
11	3	41	15	1	23	10	5	2	10
20	0	7	6	9	0	3	1	15	5
27	8	26	8	1	1	11	2	4	4
8	41	29	18	8	5	2	10	1	0
5	36	3	4	9	5	10	8	0	7

Opgave: Maak een frequentietabel

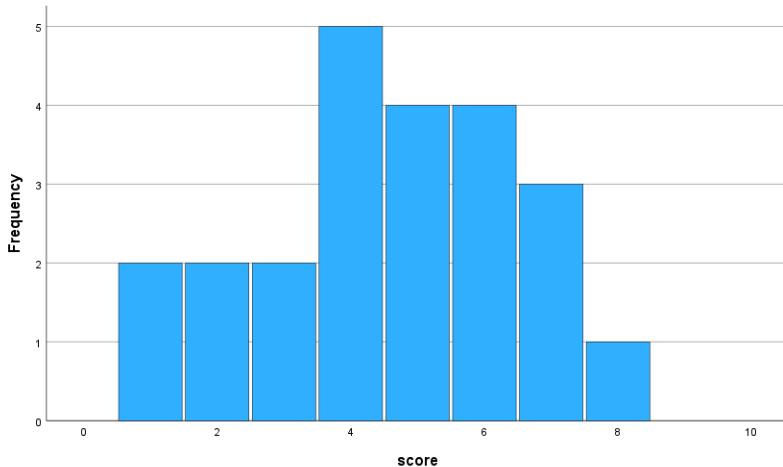
We groeperen eerst alle statistische eenheden met dezelfde waarde samen: zonder eerdere arrestaties, één eerdere arrestatie, twee eerdere arrestaties, enz. tot we alle scores in data hebben gegroepeerd. Vervolgens ordenen we deze scores in volgorde van grootte. Door op deze manier naar de data te kijken, kunnen we een idee krijgen van de aard van de verdeling van de scores. In de praktijk is het maken van een frequentieverdeling meestal de eerste stap die een onderzoeker neemt bij het analyseren van de resultaten van een onderzoek.

- a. Hoeveel bedraagt totale n ?
- b. Wat is de modus?
- c. Wat is de minst frequente score?
- d. Hoeveel respondenten hebben de score van 8?
 - a. In absolute aantallen
 - b. In relatieve frequenties
- e. Hoeveel respondenten hebben de score van 26?
 - a. In absolute aantallen
 - b. In relatieve frequenties

Tabel*Frequentieverdeling van eerdere arrestaties van 100 bekende overtreders*

Waarde	Absolute frequentie	Relatieve frequentie	Cumulatieve frequentie

2. Bekijk onderstaande histogram en beantwoord de vragen.



- Wat is een histogram?
- Hoeveel bedraagt totale n ?
- Wat is de modus of de meest frequente score?
- Wat is de minst frequente score?
- Hoeveel respondenten hebben de score van 5?
- Hoeveel respondenten hebben de score van 2?

3. Aan studenten uit de faculteit Recht en Criminologie wordt gevraagd hoe vaak ze hard-drugs gebruiken. Onderstaande tabel geeft de absolute frequenties weer. Vul de tabel aan met n, cumulatieve frequentie, proportie en cumulatieve proportie.

Categorie	Frequentie	Cumulatieve frequentie	Proportie	Cumulatieve proportie
Nooit	276			
Zelden	8			
Af en toe	4			
Dikwijls	2			
Zeer vaak	1			
	n=			

- Hoeveel studenten nemen dikwijls hard-drugs?
- Hoeveel procent is dat?
- Hoeveel procent van de studenten neemt zelden of nooit hard-drugs?
- Welke proportie van de studenten neemt minstens “dikwijls” hard-drugs?
- Stel dat de 100 mensen die niet hebben geantwoord, dat niet gedaan hebben omdat ze niet willen toegeven dat ze “dikwijls” hard-drugs gebruiken. Wat gebeurt er met de verschillende waarden als die mensen wel hadden geantwoord?

4. Een groep van 20 gevangenens in een bepaald cellenblok werd getest op hun kennis van de regels van de instelling. De scores (totaal 70 punten) waren als volgt:

31	28	27	19	18	18	41	0	30	27
27	36	41	64	27	39	20	28	35	30

- Bereken de range of variatiebreedte.
- Verwijder de laagste en de hoogste score en bereken de variatiebreedte van de resterende scores.
- Hoe verklaar je het verschil tussen waarden van deze twee maten van spreiding?

5. Hieronder vind je drie frequentietabellen. Bepaal voor elke frequentietabel *de best passende maten* van centraliteit en maten van spreiding.

Wettelijke vertegenwoordiging voor witteboordencriminaliteit	Frequentie
Geen	20
Rechtshulp	26
Door rechtbank aangeduid vertegenwoordiger	92
Openbare aanklager	153
Privé-advocaat	380
Totaal	671

Attitude van studenten over openbare dronkenschap	Frequentie
Helemaal niet erg	73
niet erg	94
neutraal	27
erg	65
Heel erg	22
Totaal	281

Aantal eerdere arrestaties	Frequentie
0	4
1	1
2	2
4	3
5	3
7	4
8	2
10	1
Totaal	20

6. De toeristische dienst van de Vlaamse kust wilt weten hoe tevreden de toeristen zijn over hun daguitstapje naar de kust. Één van de vragen luidt: "Was u tevreden met de bediening in de horecazaken ?" De verdeling van de toeristen over de antwoorden vind je terug in onderstaande frequentietabel.

ANTWOORD	AANTAL TOERISTEN
Zeer ontevreden	33
Ontevreden	84
Noch tevreden, noch ontevreden	102
Tevreden	63
Zeer tevreden	48

- a. Vul de tabel verder aan met de cumulatieve absolute frequenties, relatieve frequenties en cumulatieve relatieve frequenties.
- b. Wat is het meetniveau van de variabele ?
- c. Bereken alle relevante maten van centraliteit. (hoe weet je welke maat van centraliteit het meest relevant is ?)
- d. Bereken alle relevante maten van spreiding.

7. We vroegen aan 30 vrouwen en 30 mannen hoeveel minuten zij op een gewone avond in de week studeren. Hieronder volgen hun antwoorden:

VROUWEN						MANNEN				
180	120	180	360	240		90	120	30	90	200
120	180	120	240	170		90	45	30	120	75
150	120	180	180	150		150	120	60	240	300
200	150	180	150	180		240	60	120	60	30
120	60	120	180	180		30	230	120	95	150
90	240	180	115	120		0	200	120	120	180

- Bereken voor beide groepen de maten van centraliteit.

	VROUWEN	MANNEN
Modus		
Mediaan		
Gemiddelde		

- Beschrijf voor beide groepen de maten van centraliteit.

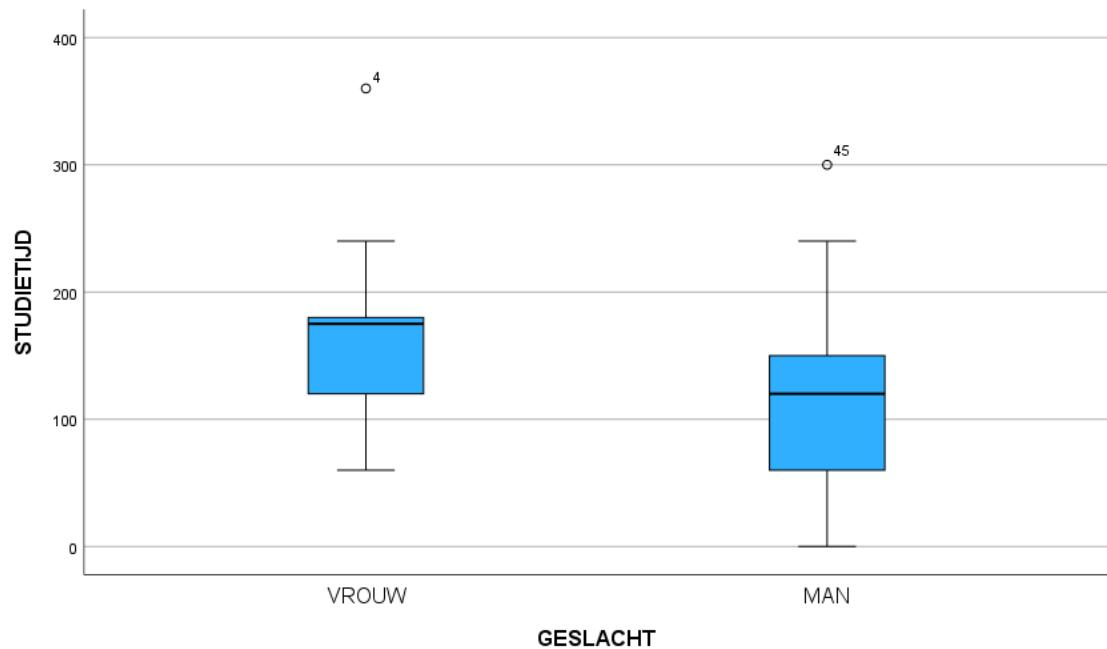
Modus

Mediaan

Gemiddelde

Hieronder vind je de visuele weergave van de verdeling van de variabele 'studietijd' voor de vrouwen en voor de mannen aan de hand van twee boxplots.

Beschrijf de vorm van de verdeling.



- Als je de boxplot van mannen en vrouwen met elkaar vergelijkt, wat kun je zeggen over:
 - Het centrum van de verdeling?
 - De spreiding van de variabele?

8. Van een groep van 20 mensen is bekend hoeveel uren ze per week studeren:

24 36 35 28 24 28 24 36 32 36 40 38 36 34 40 36 32 36 40 36

- a. Bereken de modus, mediaan en het gemiddelde.
- b. Bereken alle behandelde spreidingsmaten.
- c. Welke parameter van centraliteit en spreiding zou jij kiezen om deze variabele te beschrijven en waarom ?

9. Van de groep van 20 mensen uit vorige vraag is niet alleen het aantal uren bekend dat zij per week studeren maar ook hun geslacht:

Vrouwen : 24 36 35 28 24 28 24 36 32 36

Mannen : 40 38 36 34 40 36 32 36 40 36

- a. Bereken het gemiddelde en de standaardafwijking voor elke groep.
- b. Wat kun je zeggen over de verschillen tussen mannen en vrouwen in studieduur op basis van jouw berekeningen ?

10. De trainer van het universitaire zwemteam onderzoekt of er een effect is van het aantal keer trainen per week op hoe lang ze kunnen watertrappelen. De gegevens die hij verzamelde, vind je terug in tabel 2(te vinden onder documenten). Bereken de meest relevante maat van spreiding voor de afhankelijke variabele

Aantal minuten watertrappelen	Aantal keer trainen per week
0	3
4	1
8	2
3	1
5	0
6	0
1	1
2	0
7	1

Maak uw keuze

- 7,5
- 2,74
- 2,8
- 7,85

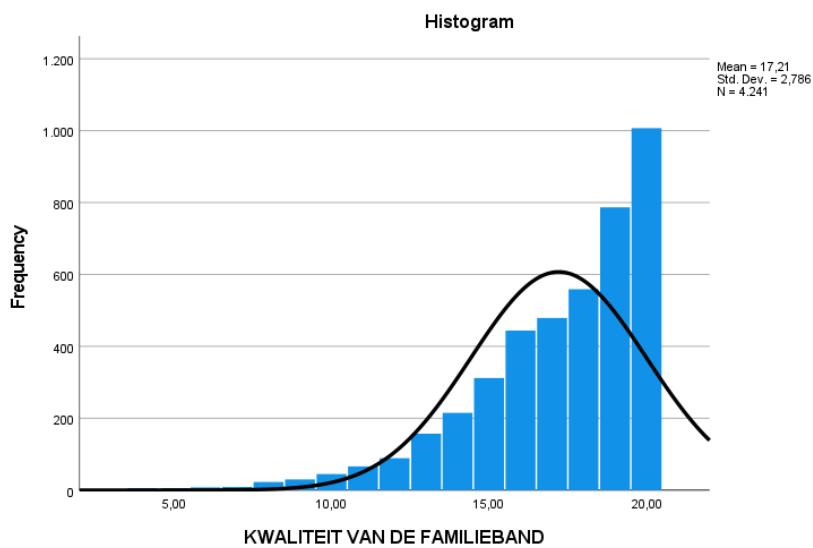
11. Variantie in bedragen van borgtocht voor een steekproef van 15 arrestanten

15 arrestanten	Bedrag van borgtocht (€)
1	500
2	1000
3	1000
4	1000
5	1200
6	1500
7	2500
8	2500
9	2500
10	2750
11	5000
12	5000
13	5000
14	7500
15	10000

Opdracht:

- Hoeveel bedraagt de standaardafwijking van borgtocht-bedragen in deze steekproef van 15 gearresteerde personen?
- Wat betekent het als de standaardafwijking gelijk is aan nul? Wat zou dit betekenen voor de scores van de variabele 'bedrag van borgtocht'?
- In welke eenheid is de standaardafwijking uitgedrukt?
- Zijn outliers een probleem voor de interpretatie van de standaardafwijking? Leg uit.

12. Bekijk onderstaande histogram en beantwoord de vragen.



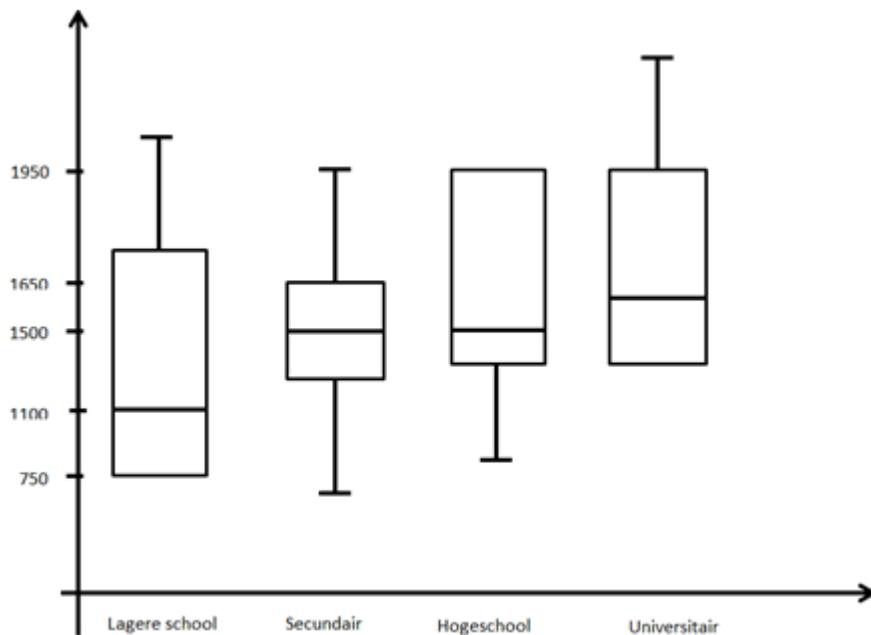
Noot: Hoge waarden verwijzen naar een goede kwaliteit van de familieband.

- a. Wat geeft een histogram weer ?

- b. Beschrijf de scheefheid van deze verdeling.

- c. Waar bevindt zich het gemiddelde ten opzichte van de mediaan in deze verdeling ?

13. De verdeling van het maandloon wordt voor vier opleidingscategorieën weergegeven in figuur 1. Welke uitspraak is JUIST?



- Personen met een hogeschool diploma verdienen gemiddeld 1500 euro per maand; personen met een diploma van de lagere school verdienen gemiddeld 1100 euro per maand.
- 50% van de respondenten met een hogeschool diploma verdient tussen de 1500 en 1950 euro per maand; eenzelfde percentage van de respondenten met een diploma van de lagere school verdient tussen de 750 en 1100 euro per maand.
- Het maandloon varieert het meest bij personen met een secundair diploma: daar is de spreiding het grootst.
- Het percentage mensen met een secundair diploma dat tussen de 1500 en 1650 euro per maand verdienen, is kleiner dan het percentage mensen met een diploma hogeschool dat tussen 1500 en 1950 euro per maand verdienen

14. Het gemiddelde examenresultaat op het statistiekexamen bedraagt 9.3 op 20 met een variantie van 2.4.

Stel dat er een fout in het examen zit en dat de studenten elk 2 punten extra verdienen. Wat gebeurt er met het GEMIDDELDE ? Wat gebeurt er met de VARIANTIE ?

- Het gemiddelde en de variantie veranderen niet.
- Het gemiddelde wordt groter en de variantie wordt kleiner.
- Het gemiddelde wordt groter en de variantie blijft dezelfde.
- Het gemiddelde en de variantie worden allebei groter.

15. Het gemiddelde examenresultaat op het examen statistiek bedraagt 9.3 op 20 met een variantie van 2.4.

Stel dat de bekendmaking van de resultaten op 100 gebeurt. We moeten alle aangepaste scores vermenigvuldigen met een factor 5. Wat gebeurt er nu met het GEMIDDELDE ? En met de VARIANTIE ?

- Het gemiddelde en de variantie worden allebei 5x groter.
- Het gemiddelde wordt 5x groter en de variantie wordt 5x kleiner.
- Het gemiddelde wordt groter en de variantie blijft dezelfde.
- Het gemiddelde wordt 5x groter en de variantie wordt 5^2 keer groter

16. Bob tracht de vijf-getallen samenvatting te berekenen voor zijn examenscores. Zijn resultaten zijn als volgt:

Minimum = 30

Maximum = 90

Q1 = 50

Q3 = 80

Mediaan = 85

Vraag: Wat zit fout aan de vijf-getallen samenvatting van Bob?

17. Statistiek in de praktijk: vergelijking tussen mediaan en gemiddelde.

- De algemene regel is dat het gemiddelde de beste maat is voor de centrale tendens van een metrische variabele...

Leg uit.

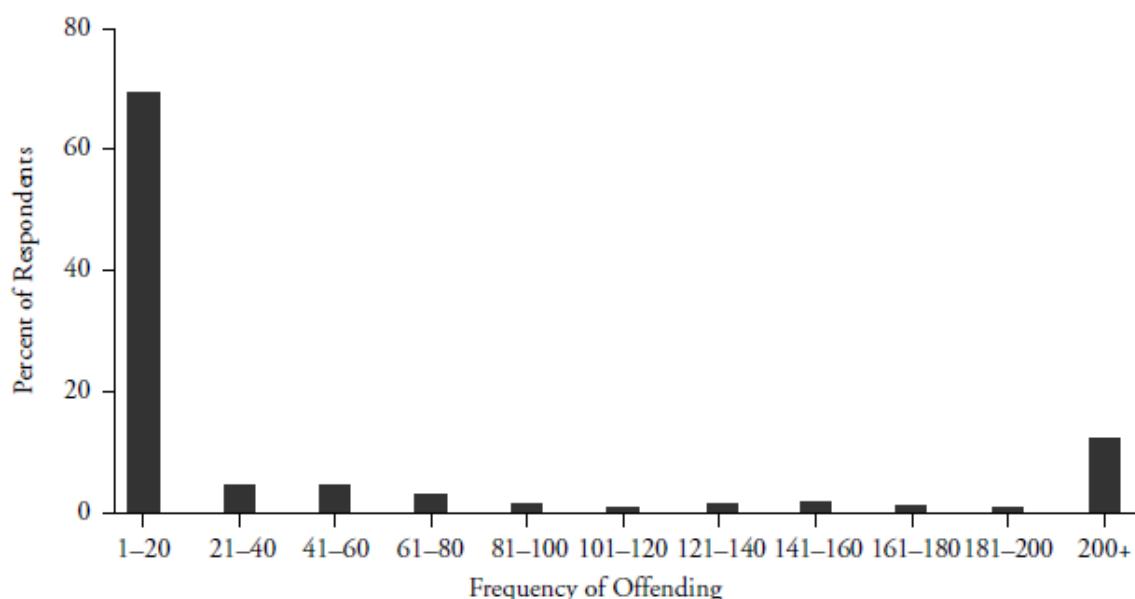
- ... maar als de verdeling van een variabele sterk scheef is (skewed), dan geeft de median een betere schatting van de centrale tendens.

- Wat betekent: "skewed"(scheef verdeeld)?

- Wat is een 'negatieve skew'?

- Wat is een 'positieve skew'?

- Bekijk onderstaande verdeling van zelfgerapporteerde criminaliteit van veroordeelde personen. Op de X-as staat het aantal eerdere delicten. Het gemiddelde in deze verdeling is 175 eerder gepleegde delicten.



Beschrijf in eigen woorden de vorm van deze verdeling. Welke maat van centraliteit zou jij verkiezen om de centrale tendens te beschrijven?

18. Een klein boekhouder kantoor betaalt aan ieder van zijn vijf administratieve medewerkers €25.000, aan twee junior-boekhouders ieder €60.000, en aan de eigenaar van de firma €225.000.

- Hoe groot is het door deze firma betaalde gemiddelde salaris?
- Hoeveel werknemers ontvangen minder dan het gemiddelde?
- Hoe hoog ligt het mediale salaris?
- Dit jaar geeft de firma enkel aan de eigenaar een salarisverhoging tot €455.000. Welk effect heeft deze wijziging op het gemiddelde?
- Welk effect is er voor de mediaan?

19. Een onderzoeker controleerde de reactietijden van de politie op tien noodoproepen telefoongesprekken. In de onderstaande gegevens staat het aantal minuten dat verstrekken tussen het einde van het telefoongesprek en het tijdstip waarop de politie arriveerde:

24 26 14 27 198 22 27 17 19 29

- a. Bereken de modus, de mediaan en het gemiddelde
- b. Welke van deze centraliteitsmaten is het meest geschikt? Verklaar jouw antwoord.

20. In een paper over extreem weer schrijft een student het volgende:

"In de meeste staten van de US komen soms orkanen voor. Wanneer een orkaan toeslaat, zijn de verliezen catastrofaal. De gemiddelde jaarlijkse verliezen zijn geen zinvolle maat voor schade voor deze zeldzame, potentieel catastrofale gebeurtenissen"

Leg uit waarom student gelijk heeft.

21. UITSPRAKEN

Uitspraak	Juist	Fout
Als de skewness nul bedraagt, wijkt de verdeling van een kenmerk niet af van de normaalverdeling.		
De som van alle gekwadrateerde afwijkingen van het gemiddelde is nul.		
De mediaan is het middelpunt van de verdeling, namelijk de waarde van de variabele die toelaat de waarnemingen in twee gelijke delen op te splitsen.		
Bij de berekening van het rekenkundig gemiddelde worden de waarden van alle geldige waarnemingen gebruikt.		
Het rekenkundig gemiddelde kan enkel worden berekend op interval niveau.		
De mediaan kan zowel berekend worden voor een variabele gemeten op het nominaal als op het ordinaal meetniveau.		
Het 50ste percentiel is gelijk aan de mediaan of aan het vijfde deciel.		
Bij nominale data is er geen sprake van rangorde, de waarde die een variabele kan aannemen heeft alleen de betekenis van een naam, een categorie. Onderzoekseenheden worden onderverdeeld in categorieën. Daarbij gelden twee belangrijke voorwaarden: exclusiviteit en exhaustiviteit. LEG UIT. Exclusiviteit =		
Exhaustiviteit =		

EXTRA OEFENINGEN

1. Binnen de Ugent wordt aan een nieuw griepvaccin gewerkt. Tijdens de tweede fase van het onderzoek naar de doeltreffendheid van dit vaccin wordt het middel uitgetest op een beperkt aantal proefpersonen. De Ugent vond enkele studenten bereid om aan het onderzoek deel te nemen. Van die proefpersonen werd de voornaam en de leeftijd genoteerd.

Voornaam	Leeftijd
Bart	23
Leen	27
Marijke	29
Johan	31
Dirk	36
Marcel	44
Sofie	28
Anne-Marie	33
Anneleen	22
Hans	27
Jeroen	24
Sonja	38
Herman	46
Tom	21
Mieke	30
Luc	37
Sarah	38
Patrick	45
Rita	48
Nathalie	26

- Hoe groot is N ?
- Wat is de modus?
- Bereken het rekenkundig gemiddelde.
- Bereken de mediaan.
- Hoeveel bedraagt de standaardafwijking?
- Is de verdeling van dit kenmerk symmetrisch, positief scheef of negatief scheef?

2. Gent werd voor de gelegenheid van een criminologische studie verdeeld in 4 gebieden: Noord, Zuid, Oost en West. Een student vergelijkt het aandeel van de inbraken van elk gebied. De relatieve frequentieverdeling van deze vier gebieden vind je in onderstaande tabel.

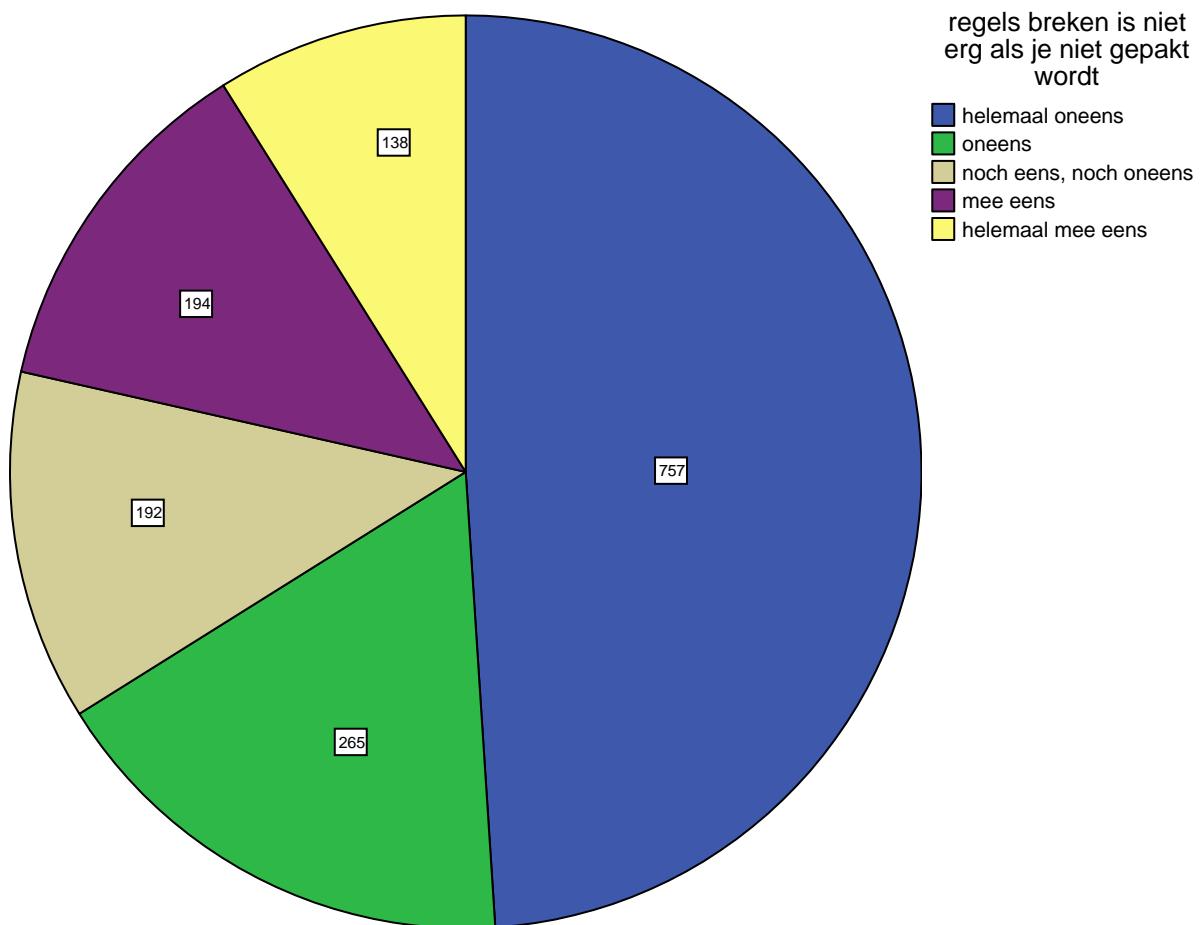
Gebied	Relatieve frequentie
Gent-Oost	0,18
Gent-Noord	0,22
Gent-Zuid	0,40
Gent-West	

- a. Wat is de relatieve frequentie van het gebied Gent-West?
- b. Het absolute aantal inbraken bedraagt 800 door de politie geregistreerde gevallen.
Wat is de absolute frequentie van Gent-Zuid?
3. Een vragenlijst gaf als antwoorden 68 maal ja, 42 maal nee en 10 maal weigering. Hieronder volgen twee uitspraken. Zijn deze juist of fout?
- N bedraagt $68+42+10$.
 - Valid N bedraagt $68 + 42$.
4. Hieronder vind je de ruwe data van het aantal e-mails dat een professor effectief moet beantwoorden per dag in een kleine faculteit gedurende een periode van twee weken.

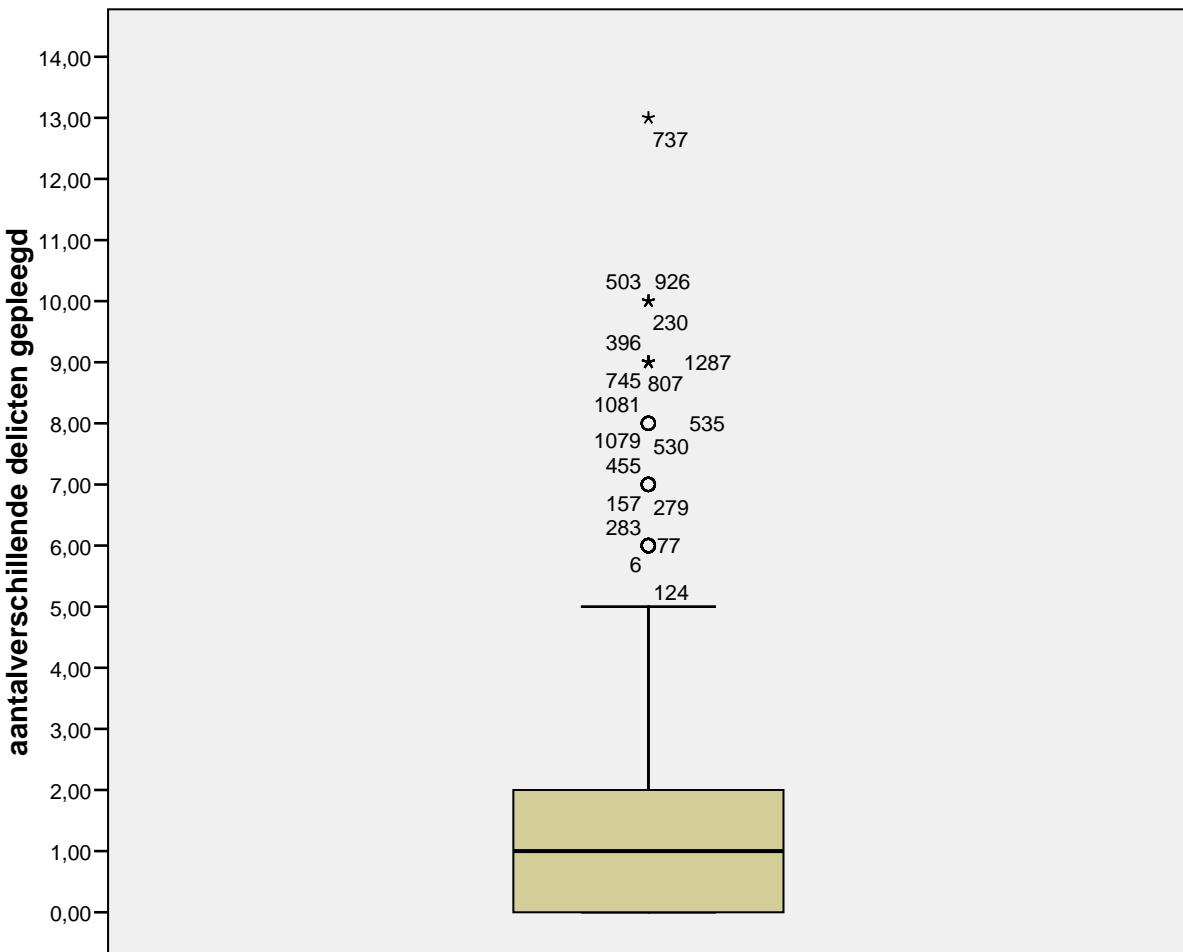
18 10 15 13 17 15 12 15 18 16

- Bereken de modus, de mediaan, het gemiddelde
- Bereken de standaardafwijking en de variatiecoëfficiënt.

5. Hieronder zie je een taartdiagram. De statisticus heeft echter de absolute waarden weergegeven.
- Geef de percentages.
 - Wat is de modus?
 - Bereken ook de variatieratio, de Index van diversiteit en de spreidingsmaat ‘d’ en interpreteer!



6. Hieronder zie je een box-plot. Bestudeer deze goed en beantwoord de uitspraken erover.



Uitspraak	Juist	Fout
Het eerste deciel valt samen met het eerste kwartiel.		
Er zijn twaalf outliers.		
Het derde kwartiel komt overeen met de waarde twee.		
De variabele is negatief asymmetrisch verdeeld.		

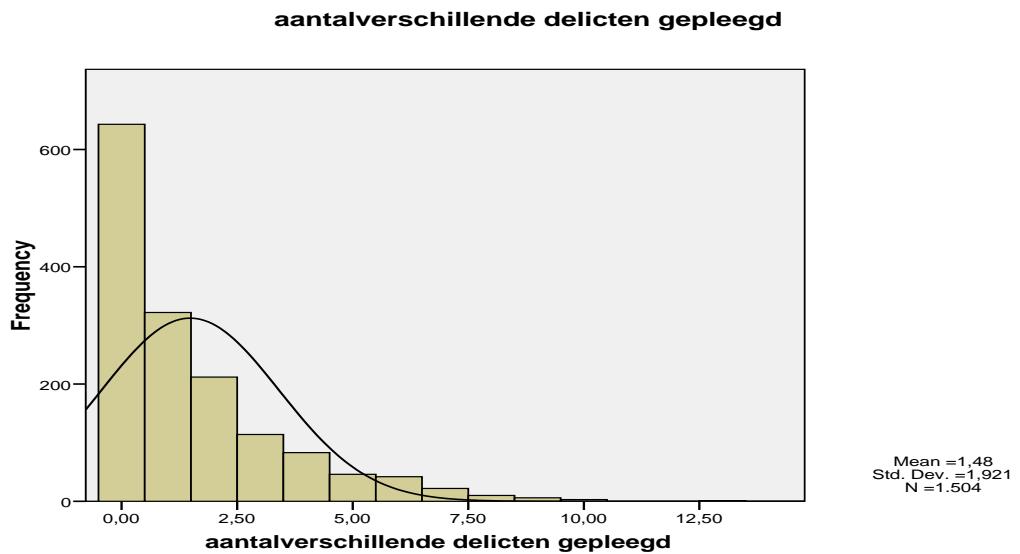
7. Bestudeer onderstaande frequentietabel aandachtig. Wat valt op het eerste zicht al op?

aantalverschillende delicten gepleegd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	643	41,4	42,8	42,8
	1,00	322	20,7	21,4	64,2
	2,00	212	13,6	14,1	78,3
	3,00	114	7,3	7,6	85,8
	4,00	83	5,3	5,5	91,4
	5,00	46	3,0	3,1	94,4
	6,00	42	2,7	2,8	97,2
	7,00	22	1,4	1,5	98,7
	8,00	10	,6	,7	99,3
	9,00	6	,4	,4	99,7
	10,00	3	,2	,2	99,9
	13,00	1	,1	,1	100,0
	Total	1504	96,8	100,0	
Missing	System	50	3,2		
	Total	1554	100,0		

Statistics

	aantalverschillende delicten gepleegd
N	1504
Valid	
Missing	50
Skewness	1,679
Std. Error of Skewness	,063
Kurtosis	3,006
Std. Error of Kurtosis	,126



- Beschrijf de scheefheid en de platheid van de variabele *aantal delicten gepleegd*
 - Wijkt de verdeling van deze variabele af van de normale verdeling ? Zo ja, hoe ?
- Is deze verdeling met andere woorden positief of negatief asymmetrisch ?

8. Hieronder vind je de SPSS output van univariate beschrijvende statistieken. Bespreek de parameters waarvan je weet dat je deze mag bespreken op basis van het meetniveau. Opgepast, er zitten een aantal addertjes onder het gras!

Statistics

		hoeveel van je vrienden hebben al eens gespijbeld?	regels breken is niet erg als je niet gepakt wordt	lid jeugdbende
N	Valid	1546	1546	1554
	Missing	8	8	0
Mean		,3234	2,1533	,06
Median		,0000	2,0000	,00
Mode		,00	1,00	0
Std. Deviation		,51783	1,37727	,247
Variance		,268	1,897	,061
Skewness		1,446	,833	3,533
Std. Error of Skewness		,062	,062	,062
Kurtosis		2,148	-,702	10,493
Std. Error of Kurtosis		,124	,124	,124
Range		3,00	4,00	1
Minimum		,00	1,00	0
Maximum		3,00	5,00	1
Percentiles	25	,0000	1,0000	,00
	50	,0000	2,0000	,00
	75	1,0000	3,0000	,00

- a. Variabele *hoeveel van je vrienden hebben al eens gespijbeld*
- b. Variabele *regels breken is niet erg als je niet gepakt wordt*
- c. Variabele *lid jeugdbende*

9. Vul de ontbrekende output verder aan.

Onderstaande tabellen uit een onderzoek van de Vlaamse regering geven de resultaten weer van een peiling naar het vertrouwen van Vlamingen ouder dan 18jaar in de politie en het gerecht. De respondenten zijn een steekproef van de Vlaamse bevolking. Beantwoord de volgende vragen:

VERTROUWEN POLITIE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ZEER WEINIG	56	3,8	3,8	3,8
	WEINIG	191	12,9	13,0	16,8
	NIET VEEL EN NIET WEINIG	528	35,7	36,0	52,8
	VEEL	609	41,2	41,5	94,3
	ZEER VEEL	84	5,7	5,7	100,0
	Total	1468	99,4	100,0	
Missing	Weet niet	7	,5		
	Geen	2	,1		
	Total	9	,6		
Total		1477	100,0		

Vertrouwen Gerecht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ZEER WEINIG	154	10,4	10,8
	WEINIG	433	29,3	30,3
	NIET VEEL EN NIET WEINIG	512	34,7	35,8
	VEEL	299	20,2	20,9
	ZEER VEEL	32	2,2	2,2
	Total	1430	96,8	100,0	
Missing	Weet niet	44	3,0		
	Geen antwoord	3	,2		
	Total	47	3,2		
Total		1477	100,0		

- a. Wat zijn de statistische eenheden ?
- b. Wat zijn de variabelen en wat is het meetniveau ?
- c. Hoeveel respondenten hebben deze vragen ingevuld (bespreek per vraag)?
- d. Wat is het verschil tussen het 'percent' en het 'valid percent'?
- e. Vul op de stippellijnen de cumulatieve percentages aan? Wat geven de cumulatieve percentages weer?
- f. Welke zijn hier de relevante parameters van centraliteit?
- g. Bespreek deze relevante parameters van centraliteit.
- h. Hebben de respondenten over het algemeen meer vertrouwen in de politie of in het gerecht?
- i. Op welke manier zou je deze frequentietabellen grafisch voorstellen

HOOFDSTUK 3

TOEPASSINGEN OP KANSREKENEN



HOOFDSTUK III

Toepassingen op kansrekenen

1. Inleiding

In dit deel worden oefeningen op het theoretische deel over kansrekenen aangeboden. Van dit deel wordt verwacht dat je de basisprincipes achter het kansrekenen kent en de centrale begrippen goed verstaat. De oefeningen zijn representatief voor de examenvragen die afkomstig zijn uit dit deel.

2. Zijn de volgende uitspraken juist of fout?

Uitspraak	Juist	Fout
Een kans is steeds een propotie.		
De kansdefinitie van Laplace is een objectieve kansdefinitie.		
De experimentele wet zegt dat naarmate het aantal herhalingen van een toevalsproces zich voordoet, de kansen van de elementen van S gaan stabiliseren.		
Een subjectieve kans is een eigen inschatting van een gebeurtenis.		
Bij de algemene somregel zijn we geïnteresseerd in de kansen dat twee gebeurtenissen afzonderlijk plaatsvinden, maar niet gezamenlijk.		
Bij de speciale somregel hebben A en B niks met elkaar te maken.		
De voorwaardelijke kans is de kans gegeven A onder de conditie B en wordt berekend aan de hand van de algemene productregel.		
De speciale productregel passen we toe wanneer we kansen willen berekenen van twee gebeurtenissen die niks met elkaar te maken hebben.		
Het aantal permutaties is het aantal mogelijkheden dat je een aantal verschillende objecten tegenover elkaar kan plaatsen.		
De verwachte waarde is de som van alle uitkomsten vermenigvuldigd met de kans op elke uitkomst.		
$P(A \text{ of } B) = P(A) + P(B) - P(A \text{ en } B)$ is een voorbeeld van de bijzondere somregel.		
$P(A \text{ of } B) = P(A) + P(B)$ is een voorbeeld van de algemene somregel.		
$P(B A) = P(A \text{ en } B) / P(A)$ is een voorbeeld van de bijzondere productregel.		
$P(B) * P(A)$ is een voorbeeld van de algemene productregel.		

3. Reken volgende kansen eigenhandig uit

- Bereken het aantal combinaties uit volgende objecten: A, A, B, B, B, C, C, C

- Bij de lotto worden iedere week zes lottogetallen getrokken door achter elkaar zes balletjes uit een machine te laten rollen. Op iedere balletje staat een getal. De balletjes die er uit zijn gerold, worden niet terug gestopt. De volgorde van de balletjes is niet belangrijk. Er zijn 41 balletjes en er worden zes balletjes getrokken. Hoeveel verschillende combinaties kunnen er voorkomen?

- Hoeveel permutaties kan je maken met 7 objecten?

- Bereken het aantal permutaties van examenvragen in een examen met 10 examenvragen.

- Bereken de kans dat 2 proefpersonen uit 5 hervallen. Je weet op voorhand dat de kans op herval .30 is.

- Bereken de kans dat 1 gevangene op 10 ontsnapt, als je weet dat de kans op ontsnapping 0.01 is.

- Wat is de kans dat 7 gevangenen op 10 ontsnappen als je weet dat de kans op ontsnapping 0.01 is?

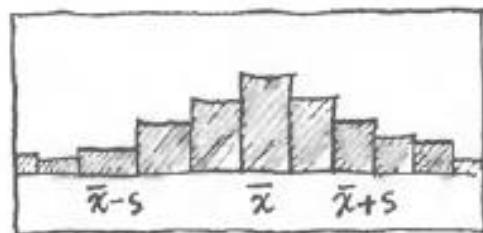
- Stel dat we met zekerheid weten dat rechters een kans hebben van 0,005 om een beklaagde onschuldig te veroordelen tot een gevangenisstraf. (Het gaat dus om een foutieve beslissing van de rechter). Als er nu 140 beklaagden voor de rechter zijn gekomen en allen waren onschuldig, wat is dan de kans dat de rechter toch een onschuldige zal veroordeeld hebben?
- Overheidsgegevens in de VS kennen aan elke moord één enkele doodsoorzaak toe. De data tonen dat de kans dat een aselect gekozen sterfgeval te wijten is aan een schotwonde, gelijk is aan 0,45 en de kans dat het sterfgeval te wijten is aan wurgscheiding, gelijk is aan 0,22. Hoe groot is de kans dat de doodsoorzaak wurgscheiding of een schotwonde was? Hoe groot is de kans dat het sterfgeval een andere oorzaak had?

HOOFDSTUK 4

DE STANDAARDNORMALE VERDELING EN DIENS EIGENSCHAPPEN

Properties of **X** and **S**

THE MEAN AND STANDARD DEVIATION ARE VERY GOOD FOR SUMMARIZING THE PROPERTIES OF FAIRLY SYMMETRICAL HISTOGRAMS WITHOUT OUTLIERS—I.E., HISTOGRAMS SHAPED LIKE MOUNDS.



HOOFDSTUK IV

STANDAARDNORMALE VERDELING

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk zijn studenten in staat toepassingen te maken op de centrale eigenschappen van normale verdelingen. Studenten kunnen vraagstukken oplossen betreffende de berekening van oppervlaktes onder de curve van de normale verdeling.

2. TE ONTHOUDEN KERNBEGRIPPEN

'68-95-99'-regel	<p>Belangrijke eigenschap van de normale verdeling:</p> <ul style="list-style-type: none"> - 68% van de scores ligt in het interval + en - 1σ rond het gemiddelde - 95% van de scores ligt in het interval + en - 1.96σ rond het gemiddelde - 99.7% van de scores ligt in het interval + en - 3σ rond het gemiddelde
Normale verdeling of Gauss-curve	Een symmetrische verdeling die gedefinieerd wordt door twee parameters: het gemiddelde en de standaarddeviatie. Dit is de bekende klokvormige frequentieverdeling.
Standaardnormale z-verdeling	<p>Bijzonder geval van normale verdeling $N(0,1)$ met gemiddelde 0 en standaardafwijking 1.</p> <p>Elke normale verdeling kan getransformeerd worden naar een standaardnormale verdeling $N(0,1)$ door elke x-score om te zetten in een z-score of gestandaardiseerde waarde.</p> <p>Voor elke normale verdeling kunnen relatieve frequenties berekend worden met behulp van de standaardnormale tabel.</p>
z-score	Gestandaardiseerde vorm van een stochastische variabele X, geeft aan hoeveel standaardafwijkingen een observatie van het gemiddelde afzit en in welke richting (positief of negatief).

3. STATISTISCHE SYMBOLEN EN FORMULES

z-score-berekening	$z = \frac{X - \bar{X}}{s}$
---------------------------	-----------------------------

4. OEFENINGEN

1. De gemiddelde leeftijd waarop langdurig heroïneverslaafden sterven is 60jaar met een standaardafwijking van 12 jaar. Beantwoord de volgende vragen:
 - 1.1. Hoeveel percent van de langdurig verslaafden sterft voor of op de dag van zijn zestigste verjaardag?
 - 1.2. Hoeveel percent van de langdurig verslaafden sterft na of op de dag van zijn zestigste verjaardag?
 - 1.3. Wat is de kans dat een langdurig verslaafde 80 wordt of ouder ?
 - 1.4. Wat is de kans dat we een langdurige verslaafde in onze steekproef aantreffen die gestorven is voor of op zijn 55^{ste} ?
 - 1.5. Hoeveel percent van de langdurige verslaafden sterft tussen zijn 54^{ste} en 67^{ste} levensjaar?
2. Stel dat de bedragen van opgelegde verkeersboetes normaal verdeeld zijn met een gemiddelde van 42.50 € en een standaardafwijking van 7.5 €.
 1. Wat is de kans dat een willekeurige overtreder een boete dient te betalen tussen de 20 € en de 25 €?
 2. Wat is de kans dat een willekeurige overtreder een boete dient te betalen lager of gelijk aan 32 €?
 3. Wat is de kans dat een willekeurige overtreder een boete dient te betalen hoger of gelijk aan 29€ ?

3. Onderzoek toont aan dat het gewicht van het volwassen brein normaal is verdeeld met een gemiddelde $\mu = 1270$ gram en een standaardafwijking die $\sigma = 137$ gram bedraagt. Hoeveel mag je brein maximum wegen om tot de centrale 60 % te behoren?

Maak uw keuze

1154,24 gram

1304,94 gram

1445,63 gram

1385,77 gram

4. Onderzoek toont aan dat de duur van menselijke zwangerschappen normaal verdeeld zijn met een gemiddelde $\mu = 266$ dagen en een standaardafwijking die $\sigma = 16$ dagen bedraagt.

a. Hoelang mag een zwangerschap maximaal duren om tot de centrale 95% te behoren?

b. Hoelang moet een zwangerschap minimaal duren om tot de centrale 95% te behoren?

5. Onderzoek toont dat de lengte van haaien perfect normaal verdeeld is met een gemiddelde $\mu = 1.5\text{m}$ en een standaardafwijking die $\sigma = 0.30\text{m}$ bedraagt.

a. Hoelang mag een haai maximaal zijn om tot de centrale 90% te behoren?

b. Hoelang moet een haai minimaal zijn om tot de centrale 90% te behoren?

6. Grottere zoogdieren hebben een langere draagtijd. De duur van de dracht bij paarden van conceptie tot geboorte varieert overeenkomstig een bij benadering normale verdeling met een gemiddelde van 336 dagen en een standaardafwijking van 3 dagen. Gebruik het 68-95-99.7- criterium voor de antwoorden op de volgende vragen.
- a. Bijna alle (99.7%) dracht bij paarden valt in welk duurbereik?
- b. Welk percentage van de dracht is langer dan 339 dagen?
7. Mannen van middelbare leeftijd zijn gevoeliger voor een hoge cholesterol dan jonge vrouwen. De cholesterolniveaus van mannen van 55 tot en 64 jaar zijn ongeveer normaal met het gemiddelde van 222 mg/dl en een standaardafwijking van 37 mg/dl.
- a. Welk percentage van deze mannen heeft een hoog cholesterol niveau (boven 240 mg/dl)?
- b. Hoeveel procent is normaal-hoog (cholesterol tussen 200 en 240 mg/dl)?

8. De Lijn Oost-Vlaanderen stelt een intern onderzoek in. Men wilt de effectiviteit van het tramverkeer in Gent meten. Eén van de resultaten wees uit dat de tijd dat tramlijn 21 erover doet om van de Zwijnaardebrug naar Melle Leeuw te rijden, perfect normaal verdeeld is volgens $N(60;5)$.
- a) Welk percentage van de trams rijdt langer over het traject dan 67,5 uur?
- b) Hoeveel minuten doen de 5% snelste trams over het traject?
- c) Welke proportie van de trams doet minder lang dan 75 minuten, maar langer dan 55 minuten over het traject?
- d) Als de Lijn op haar aankondigingen plaatst dat de tram na 57 minuten op haar eindbestemming aankomt, hoeveel procent van de trams rijdt dan op tijd?
- e) Als de Lijn in haar jaarlijkse doelstelling opgenomen heeft dat 30 % van de trams op tijd moeten aankomen, welke ritduur moet de Lijn dan op haar aankondigingen plaatsen?

HOOFDSTUK 5

BIVARIATE BESCHRIJVENDE STATISTIEK

EARTHTOPLANET.COM

Andy Zellman



"MOTHER, YOU MUSTN'T CONFLATE
CORRELATION WITH CAUSATION."

HOOFDSTUK V

BIVARIATE STATISTIEK

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk zijn studenten in staat op basis van het meetniveau van twee kenmerken te kiezen voor de meeste geschikte associatiemaat om het verband tussen twee variabelen te beschrijven. Studenten kunnen de toegepaste associatiematen correct interpreteren. Bij het toepassen van de bivariate beschrijvende statistiek is het van belang een onderscheid te maken tussen symmetrische en asymmetrische relaties en bijgevolg tussen symmetrische en asymmetrische associatiematen

2. BIVARIATE ASSOCIATIEMATEN TUSSEN CATEGORISCHE VARIABELEN**2.1. Nominale variabelen****2.1.1. Te onthouden kernbegrippen**

Chi ²	Associatiemaat voor kenmerken op het nominale niveau. Chi ² heeft een moeilijke interpretatie en geen absolute begrenzing. Wordt in hoofdzaak gebruikt in de inferentiële statistiek om na te gaan of een verband al dan niet op toeval berust. Indien Chi ² een waarde van 0 aanneemt, is er geen samenhang tussen de variabelen.
Cramer's V	Associatiemaat gebaseerd op chi ² . Varieert net als phi van 0 tot 1 maar kan ook gebruikt worden bij grotere tabellen dan 2*2. Bij een 2*2 tabel is de waarde van Cramer's V gelijk aan de waarde van phi.
Kruistabel of contingentietabel	Tabel waarin de categorieën van twee variabelen tegenover elkaar worden gezet en waarin de waargenomen frequentie van elke combinatie van categorieën vermeld staat.
Odds (ratio)	Een maat om de verhouding tussen het voorkomen van een gebeurtenis en het niet voorkomen van een gebeurtenis te beschrijven. De odds ratio is de verhouding tussen twee odds. 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.
Percentageverschil	Bivariate associatiemaat voor kenmerken op het nominale niveau. De percentages op 1 categorie van de afhankelijke variabele worden vergeleken voor de verschillende categorieën

Phi

van de onafhankelijke variabele. Het verschil wordt weergegeven in percentagepunten.

Associatiemaat gebaseerd op χ^2 . Heeft de waarde 0 bij afwezigheid van associatie en waarde 1 bij perfecte statistische associatie. Wordt gebruikt in 2×2 tabellen.

2.1.2. Statistische symbolen en formules

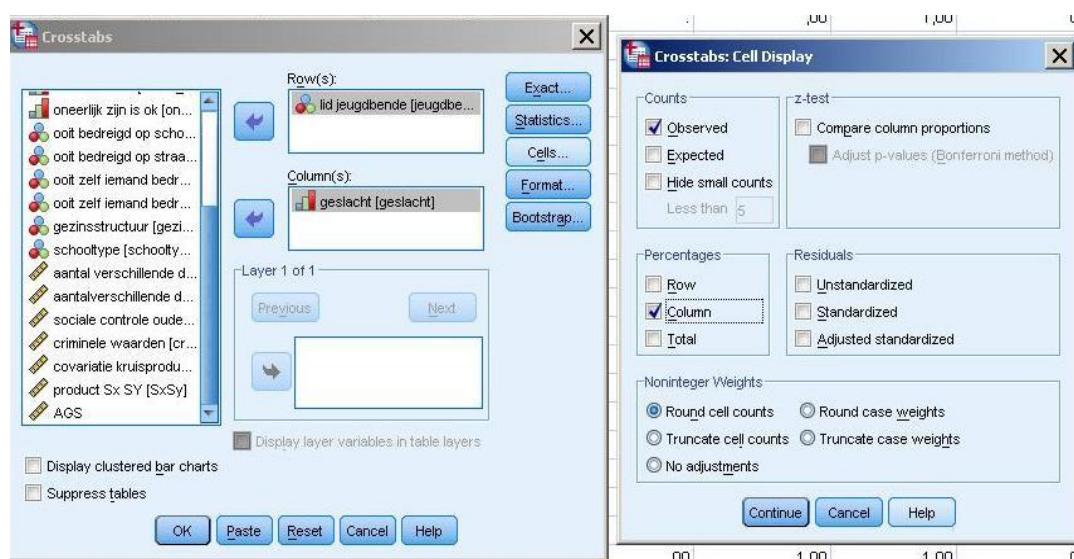
Chi²

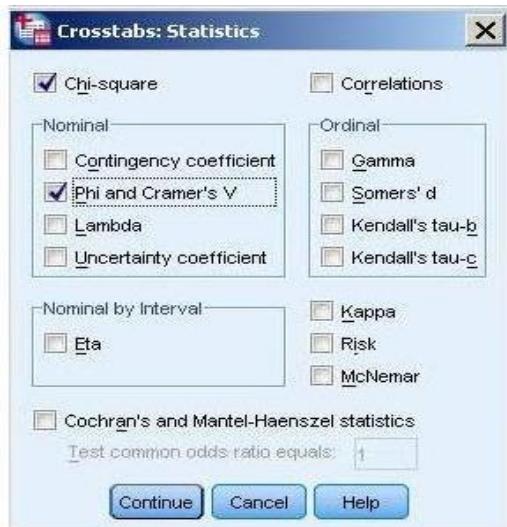
$$\chi^2 = \sum \frac{(geobserveerd - verwacht)^2}{verwacht}$$

2.1.3. Associatiematen op nominaal niveau: SPSS-output en interpretatie

Zijn jongens meer betrokken bij een jeugdbende dan meisjes?

We hebben gezien dat associatiematen op nominaal niveau worden gebruikt voor de analyse van contingentietabellen. We tonen je eerst even voor hoe je een contingentietabel in SPSS kan maken. Kies onder "Analyze" → "cross-tabs" en bepaal de afhankelijke en onafhankelijke variabele! Denk logisch na! Door te klikken op "statistics" krijgen we de associatiematen te zien. We vragen enkel de associatiematen op die passen bij het meetniveau. We maken de kruistabel door te klikken op "cells" en de juiste percentages op te vragen. Je moet verstandig kiezen tussen kolompercentages en rijpercentages. Als je de afhankelijke in de rijen plaatst, dan bereken je de kolompercentages.





Je klikt vervolgens op de associatiematen die je wil kennen. Wij kiezen voor chi-kwadraat, Phi en Cramer's V als belangrijke nominale associatiematen.

Zo ziet de output van de statistische analyse van de hierboven beschreven contingentietabel in SPSS er uit:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
lid jeugdbende * geslacht	1548	99,6%	6	,4%	1554	100,0%

lid jeugdbende * geslacht Crosstabulation

			geslacht		Total
			meisje	jongen	
			N	%	
lid jeugdbende	geen lid jeugdbende	Count	786	661	1447
		% within geslacht	96,6%	90,1%	93,5%
	lid jeugdbende	Count	28	73	101
		% within geslacht	3,4%	9,9%	6,5%
Total		Count	814	734	1548
		% within geslacht	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	26,785 ^b	1	,000		
Continuity Correction ^a	25,729	1	,000		
Likelihood Ratio	27,447	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	26,768	1	,000		
N of Valid Cases	1548				

a. Computed only for a 2x2 table

b. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 47,89.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,132	,000
	Cramer's V	,132	,000
N of Valid Cases		1548	

- a. Not assuming the null hypothesis.
- b. Using the asymptotic standard error assuming the null hypothesis.

Eerst krijgen we de informatie te zien over de observaties (respondenten) en de ontbrekende waarden. Daarna volgt de uiteindelijke kruistabel. Tot slot volgen associatiematen, de chi-kwadraat waarde en automatisch verkrijgen we de bijbehorende significantietoetsen.¹ We krijgen ook de associatiematen op nominaal niveau te zien: Phi en Cramer's V. Is er nu een verband tussen geslacht en het behoren tot een jeugdbende? We lezen de informatie af uit de kruistabel en interpreteren de resultaten. We bespreken de resultaten zoals het hoort in een rapport dat je maakt in het kader van de bachelorproef of masterproef.

We rapporteren als volgt. *“Deze bivariate analyse is gebaseerd op 1548 respondenten. Dit is de effectieve steekproefgrootte. De item-nonrespons bedraagt slechts 0.4% van het totaal aantal respondenten (1554). 6.5% (101) van de respondenten geeft aan lid te zijn van een jeugdbende. Jongens (9.9%) zijn vaker lid dan meisjes (3.4%). Dit geeft een verschil van 6.5 percentagepunten.”* De associatiemaat Cramer's V en Phi zijn hier aan elkaar gelijk omdat we een 2*2 tabel hebben. Er zijn twee rijen en twee kolommen omdat de beide variabelen bestaan uit twee categorieën. In de theoretische lessen hebben we gezien dat in zulke situaties de beide coëfficiënten gelijk zijn. Onthou dat je een keuze dient te maken bij de rapportage en dat je steeds dient te motiveren waarom je een bepaalde associatiemaat hanteert. Het verband is eerder aan de zwakke kant. Cramer's V bedraagt 0.132. Chi-kwadraat bedraagt 26.78. Dit betekent dat er een verschil is tussen de geobserveerde celfrequenties en de verwachte celfrequenties die men zou vinden indien er geen statistisch verband bestaat. Let op! Chi-kwadraat is geen zuivere maat. Had onze steekproef twee keer zo groot geweest, dan was de waarde voor chi-kwadraat ook twee keer zo groot.

In SPSS kan je trouwens zelf uitrekenen hoe de verdeling er zou uitzien indien er geen statistische associatie was geweest tussen beide kenmerken. Dit kan gedaan worden door de “expected counts” op te vragen.

¹ Op de significantietoetsen wordt nu niet verder ingegaan. We hernemen dit in het deel over de inferentiële statistiek (deel V uit deze toegepaste syllabus).

lid jeugdbende * geslacht Crosstabulation

			geslacht		Total	
			meisje	jongen		
lid jeugdbende	geen lid jeugdbende	Count	786	661	1447	
		Expected Count	760,9	686,1	1447,0	
		% within geslacht	96,6%	90,1%	93,5%	
	lid jeugdbende	Count	28	73	101	
		Expected Count	53,1	47,9	101,0	
		% within geslacht	3,4%	9,9%	6,5%	
Total		Count	814	734	1548	
		Expected Count	814,0	734,0	1548,0	
		% within geslacht	100,0%	100,0%	100,0%	

Mocht er geen statistische associatie bestaan tussen geslacht en het al of niet lid zijn van een jeugdbende, dan zouden er 760.9 meisjes en 686.1 jongens geen lid zijn van een jeugdbende en zouden er 53.1 meisjes en 47.9 jongens lid zijn van een jeugdbende. De cijfers na de komma hebben hier geen betekenis, maar zijn het resultaat van de toepassing van de formule. Wat je moet onthouden is het volgende: bij afwezigheid van een statistische relatie, zijn de conditionele frequentieverdelingen identiek. Als 6.5% van de observaties lid is van een jeugdbende, dan moeten de kolompercentages voor meisjes en jongens ook allebei 6.5% bedragen. We controleren even door de “expected counts” te delen door de kolomtotalen en we zien inderdaad dat deze beide 6.5% zijn, voor jongens is dat 47.9/734 en voor meisjes is dat 53.1/814. Chi-kwadraat is echter gevoelig aan het aantal rijen en kolommen en aan de marginale verdelingen. Daarom is het belangrijk ook de andere associatiematen te bekijken.

De odds en oddsratio moet je zelf kunnen berekenen op basis van een contingentietabel. De odds voor het lid zijn van een jeugdbende (versus geen lid zijn) bedraagt voor jongens 73/661 en voor meisjes 28/786. De odds ratio is dus $(73/661)/(28/786)$. Anders gesteld: de kans dat jongens lid zijn van een jeugdbende is 3.1 keer groter dan de kans dat meisjes lid zijn van een jeugdbende.

2.2. Ordinale variabelen

2.2.1. Te onthouden kernbegrippen

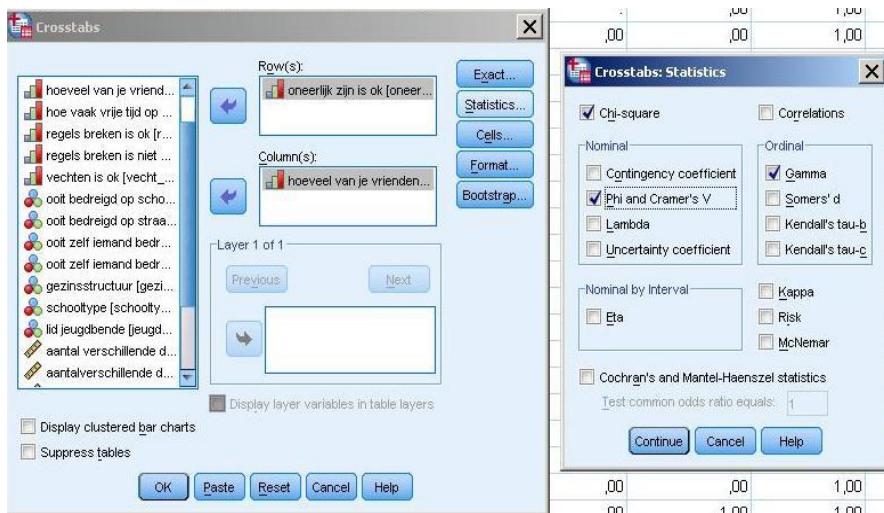
gamma	Associatiemaat voor kenmerken op het ordinale niveau. Neemt een waarde van -1 aan bij perfecte negatieve samenhang en +1 bij perfecte positieve samenhang en 0 bij afwezigheid van samenhang.
Rangcorrelatiecoëfficiënt van Kendall (Kendall's Tau-b)	Ordinal maat van samenhang en neemt waarden aan van -1 tot +1 en neemt de waarde 0 aan bij afwezigheid van een lineaire samenhang.
Rangcorrelatiecoëfficiënt van Spearman (Spearman's rho)	Associatiemaat voor kenmerken op ordinaal niveau die afgeleid is van de Pearson productmomentcorrelatiecoëfficiënt. Neemt waarden aan tussen -1 en +1 en neemt de waarde 0 aan bij afwezigheid van een lineaire samenhang

2.2.2. Associatiematen op ordinaal niveau: SPSS-output en interpretatie

Relatie tussen delinquente vrienden en oneerlijk zijn

De associatie tussen ordinale kenmerken kan gebeuren aan de hand van de associatiematen Cramer's V, Gamma en de rangcorrelatiecoëfficiënt van Spearman, nl. Spearman's rho. We gebruiken opnieuw de dataset "Oefendataset1statcrim". We gaan de samenhang tussen twee uitspraken met elkaar vergelijken. Deze zijn: "*oneerlijk zijn is ok*" (de antwoordcategorieën gaan van helemaal niet akkoord tot helemaal akkoord) en "*hoeveel van je vrienden hebben al iets gestolen?*" (de antwoordcategorieën gaan van geen enkele tot bijna allemaal). Deze variabelen zijn ordinaal want ze bestaan uit ordenbare antwoordcategorieën. Bijna allemaal is meer dan een enkele, maar de afstand daartussen is niet metrisch uit te drukken.

We tonen je eerst even hoe je de voorbeeldoefening in SPSS. We plaatsen één variabele in een rij, de andere in een kolom.



Zo ziet de output van de analyse van de contingentietabel er uit in SPSS:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
hoeveel zijn ok? * hoeveel van je vrienden hebben al iets gestolen?	1540	99,1%	14	,9%	1554	100,0%

hoeveel zijn ok? * hoeveel van je vrienden hebben al iets gestolen? Crosstabulation

		hoeveel van je vrienden hebben al iets gestolen?				Total
		een enkele	sommige	de meeste	bijna allemaal	
hoeveel zijn ok?	helemaal oneens	Count	720	41	4	0
		% within hoeveel van je vrienden hebben al iets gestolen?	51,9%	28,9%	50,0%	,0%
	oneens	Count	288	34	0	0
		% within hoeveel van je vrienden hebben al iets gestolen?	20,7%	23,9%	,0%	,0%
	noch eens, noch oneens	Count	189	24	1	1
		% within hoeveel van je vrienden hebben al iets gestolen?	13,6%	16,9%	12,5%	50,0%
mee eens	mee eens	Count	129	24	2	0
		% within hoeveel van je vrienden hebben al iets gestolen?	9,3%	16,9%	25,0%	,0%
Total	helemaal mee eens	Count	62	19	1	1
		% within hoeveel van je vrienden hebben al iets gestolen?	4,5%	13,4%	12,5%	50,0%
Total		Count	1388	142	8	2
		% within hoeveel van je vrienden hebben al iets gestolen?	100,0%	100,0%	100,0%	100,0%

Symmetric Measures

	Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Phi	,192		,000
Nominal	Cramer's V	,111		,000
Ordinal by Nominal	Gamma	,375	,055	,000
N of Valid Cases		1540	5,639	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Opnieuw krijgen we informatie over het aantal geldige observaties en het aantal ontbrekende waarden. We hebben in deze kruistabel een inhoudelijke overweging gemaakt: hoewel we de associatie tussen beide variabelen als symmetrisch zouden kunnen beschouwen, hebben we dit hier niet gedaan. We hebben ervoor geopteerd om “het oneerlijk zijn” te zien als een afhankelijke variabele, dit wil zeggen dat de waarden erop “worden beïnvloed” door een ander kenmerk, met name “het hebben van criminale vrienden”, een statement uit de theoretische criminologie, meer bepaald de theorie van Sutherland. We konden er evengoed vanuit gaan dat het oneerlijk zijn leidt tot het hebben van criminale vrienden. En in dat laatste geval hadden we niet de kolompercentages, maar de rijpercentages moeten berekenen. Laat ons nu eens trachten de diagonaal te bekijken van links boven tot rechts onder. We zien toch wel een zeker patroon, al is het niet superduidelijk. Links boven en rechts onder zijn de hoogste percentages vast te stellen: eerder lage waarden op X hangen samen met eerder lage waarden op Y, wie matig scoort op X scoort matig op Y en wie hoog scoort op X scoort ook hoog op Y. In de theoretische syllabus werd gesteld dat de analyse van consistente en inconsistente paren kan leiden tot het ontdekken van een ordinaal lineair patroon. Hoewel we de kolompercentages zouden kunnen bespreken, is het duidelijk dat dit al snel heel veel werk vereist in een grote r^*k tabel. Hierin ligt duidelijk het nut van een associatiemaat. Deze vat in één maat samen wat we hier te zien krijgen in deze 5^*4 tabel.

Phi en Cramer's V zijn symmetrische maten. Dit betekent dat geen causale richting wordt verondersteld. Zij houden ook geen rekening met de ordenbaarheid in de data. Aangezien we te maken hebben met een r^*k tabel, kijken we niet naar Phi maar naar Cramer's V. Gamma daarentegen houdt rekening met de ordenbaarheid van de data: als variabele X een hogere waarde heeft, heeft variabele Y dan ook een hogere waarde?

Hoe rapporteren we deze beschrijvende bevindingen? *“1540 respondenten werden betrokken in deze bivariate analyse. Cramer's V bedraagt 0.111. Er is een zwakke samenhang tussen de beide variabelen.”*

Gamma geeft ons meer informatie dan Cramer's V. Omdat beide variabelen ordinaal zijn, verkiezen we gamma als associatiemaat boven Cramers V. Immers, Cramer's V houdt geen rekening met de ordening in de data. Gamma zegt ons dat de associatie tussen beide kenmerken matig en positief is ($\text{Gamma} = 0.375$). Hoe meer vrienden men heeft die al eens iets gestolen hebben, hoe meer men de neiging heeft om oneerlijk te zijn. Merk op dat gamma een hogere waarde heeft dan Cramer's V.

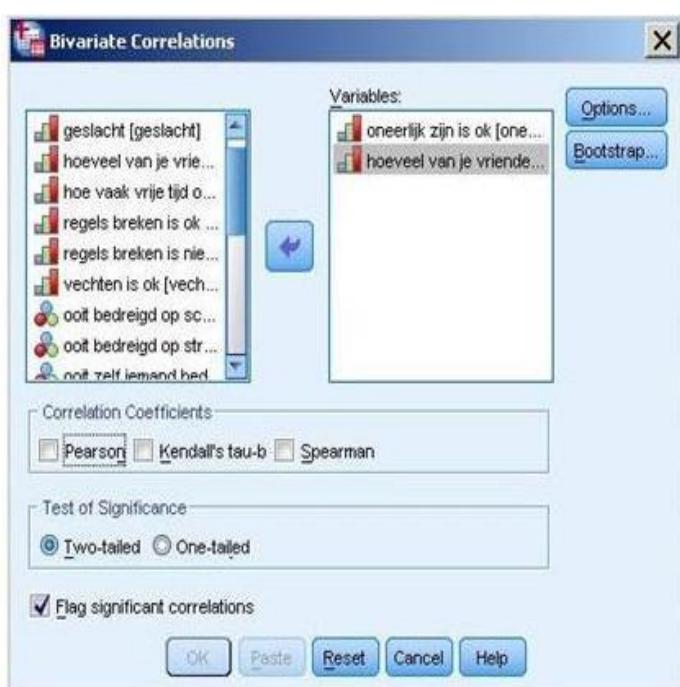
Opgelet! Gamma gaat uit van een monotone rechtlijnige samenhang tussen de beide kenmerken. In de realiteit komt het voor dat er een verband bestaat tussen beide kenmerken,

maar dat dit verband niet eenduidig rechtlijnig is. In dat verband is er sprake van een niet-lineair verband.

Let op! De relatie tussen ordinale en nominale kenmerken is gevoelig voor het kiezen van breekpunten. Stel je voor dat we deze r^*k tabel zouden herleiden tot een $2*2$ tabel door categorieën samen te gooien, dan wordt de uitkomst door dit keuzeproces beïnvloed. Als je ooit zelf een analyse maakt in het kader van je bachelorproef of masterproef, moet je hiermee rekening houden!!!!

De rangcorrelatiecoëfficiënt Spearman's Rho als alternatieve symmetrische ordinale associatiemaat

We kunnen de relatie tussen twee ordinale kenmerken tenslotte ook via Spearman's Rho berekenen. We tonen je eerst even hoe je deze analyse zelf kan uitvoeren.



In SPSS voeren we een rangcorrelatie-analyse uit als volgt:
Onder analyze → bivariate correlations en vink aan Spearman!

De output ziet er zo uit:

Correlations			
		oneerlijk zijn is ok	hoeveel van je vrienden hebben al iets gestolen?
Spearman's rho	oneerlijk zijn is ok	Correlation Coefficient Sig. (2-tailed) N	1,000 .000 1548
	hoeveel van je vrienden hebben al iets gestolen?	Correlation Coefficient Sig. (2-tailed) N	.158** .000 1540

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

Spearman's Rho, de rangcorrelatiecoëfficiënt, is afgeleid van de Pearson's product-moment correlatiecoëfficiënt voor interval- en ratio variabelen. De observaties worden eerst in een gewone rangorde (1^{ste}, 2^{de}, 3^{de}, ...) geplaatst op de beide variabelen. Daarna past men de formule voor de productmoment correlatiecoëfficiënt toe. Bij ex-aequo's moet een aangepaste formule gebruikt worden. Uit de tabel blijkt dat de rangcorrelatiecoëfficiënt 0.158 is. Dit wijst op een eerder zwakke samenhang. Mogelijks wordt deze zwakkere samenhang veroorzaakt door ex-aequo's. Wij raden aan voor de analyse van ordinale variabelen vooral Gamma te gebruiken. Deze geeft de minst vertekende samenhang tussen beide ordinale kenmerken in het geval van lineaire associatie.

3. BIVARIATE ASSOCIATIEMATEN TUSSEN METRISCHE VARIABELEN

3.1. Symmetrische associatiematen: correlatie analyse

3.1.1. Te onthouden kernbegrippen

Covariatie (SS _{Xy})	Maat van samenhang op het metrische niveau. De covariatie stelt de mate voor waarin twee kenmerken samen variëren. Het is de som van de kruisproducten van de deviatiescores van X en van Y. Ook: kruisproductensom of sum of squares
Covariantie (S _{Xy})	Maat van samenhang op het metrische niveau. Het is de som van de kruisproducten van de deviatiescores van X en Y gedeeld door n-1.
Correlatiecoëfficiënt van Pearson (R _{Xy})	Ook wel de productmomentcorrelatiecoëfficiënt van Pearson genoemd. De meest gebruikte bivariate associatiemaat van het gezamenlijk variëren of samenhang voor kenmerken van het metrische niveau. Geeft een indruk van de sterkte en de richting van de lineaire samenhang tussen X en Y. R varieert tussen -1 en +1 met 0 = afwezigheid van lineair verband. Het verband tussen de twee variabelen wordt verondersteld rechtlijnig te zijn (is het verband gekromd dan geeft de correlatiecoëfficiënt een vertekend beeld).
Lineaire samenhang of correlatie	Samenhang tussen waarden op twee variabelen in die zin dat waarden van de beide variabelen dezelfde of een tegengestelde tendens vertonen. Lineair betekent dat het gaat over 'samenhang ten opzichte van een rechte'. De correlatiecoëfficiënt is een maat voor de sterkte van de lineaire samenhang tussen X en Y.

Scatterplot of puntenwolk of spreidingsdiagram	Verzameling van alle elementen uit de steekproef waarbij elk punt (coördinaat) een statistische eenheid weergeeft met informatie op de X-variabele en de Y-variabele.
Symmetrische associatiemaat	Associatiemaat die de sterkte van een symmetrisch verband tussen variabelen weergeeft.
Symmetrische relatie	Geen expliciete veronderstelling over de causale relatie tussen variabelen. Er wordt enkel nagegaan in welke mate er samenhang bestaat tussen de variabelen. Deze samenhang geldt in beide richtingen. Vb. Als hoge waarden op A samenhang met hoge waarden op B, dan gaan hoge waarden op B ook samen met hoge waarden op A.

3.1.2. Statistische symbolen en formules

Covariatie of kruisproductensom of sum of squares (SS _{xy})	$SS_{xy} = \sum (X - \bar{X})(Y - \bar{Y})$
Covariantie (S _{xy})	$S_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1}$
Correlatiecoëfficiënt van Pearson (R _{xy})	$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$

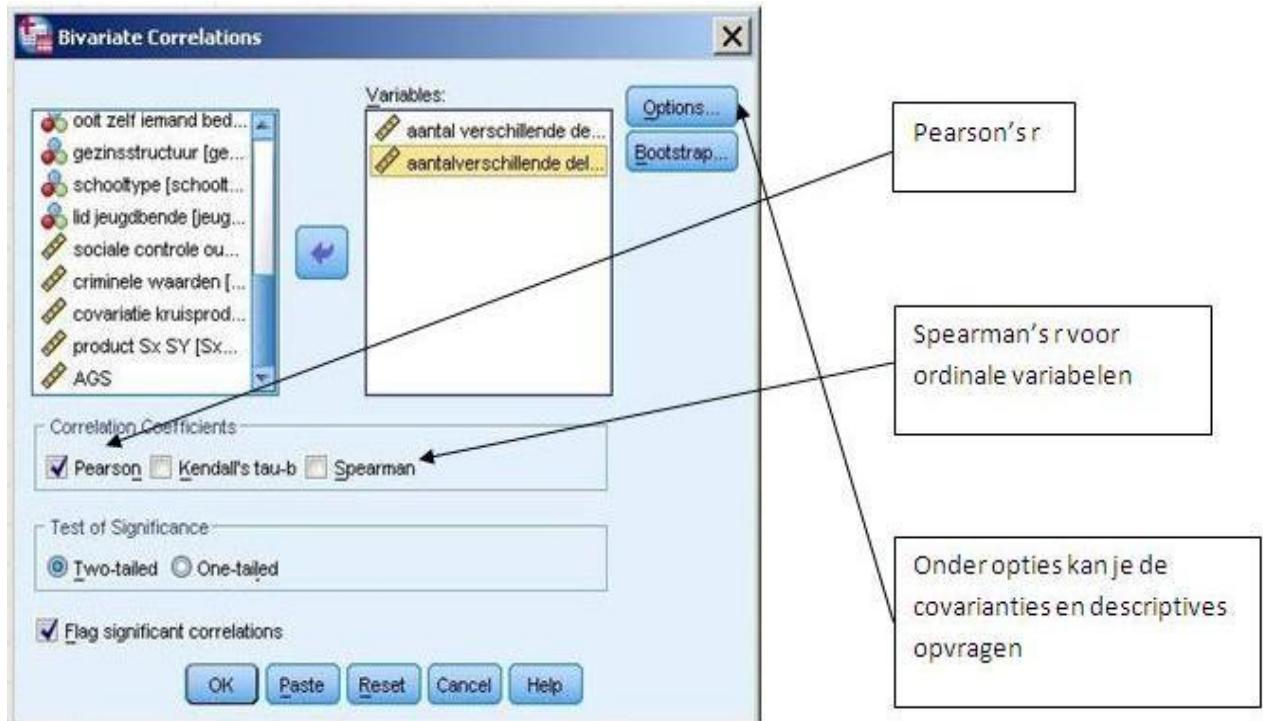
3.1.3. Symmetrische associatiematen op metrisch niveau: SPSS-output en interpretatie

Is er een samenhang tussen daderschap en slachtofferschap ?

Symmetrische associatiematen voor kenmerken gemeten op metrisch niveau zijn de covariatie, de covariantie en de correlatie. We geven het voorbeeld hoe je de correlatiecoëfficient uitrekent tussen twee metrische kenmerken. We gebruiken het aantal keer dat men slachtoffer is geweest en dader is geweest als voorbeeld. Beide concepten zijn gemeten op het metrische niveau. Het berekenen van de correlatiecoëfficiënt van Pearson gebeurt aan de hand van de formule:

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

SPSS geeft de waarde van de correlatiecoëfficiënt zelf via de procedure "correlations"



De output van een correlatie-analyse ziet er als volgt uit:

		aantal verschillende delicten slachtoffer	aantalverschillende delicten gepleegd
aantal verschillende delicten slachtoffer	Pearson Correlation Sig. (2-tailed) N	1 1491	,444** ,000 1453
aantalverschillende delicten gepleegd	Pearson Correlation Sig. (2-tailed) N	,444** ,000 1453	1 1504

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

We geven enkele richtlijnen met betrekking tot de sterke van het gevonden verband:

Deze associatiemaat gaat van -1 tot +1 en is 0 bij afwezigheid van een lineair verband. *

- waarde tussen 0 en 0.10 (in absolute waarden): verwaarloosbaar verband
- waarde tussen 0.10-0.30 (in absolute waarden) : zwak tot matig bivariaat verband
- waarde tussen 0.30-0.60 (in absolute waarden): matig tot sterk bivariaat verband
- waarde hoger dan 0.60 (in absolute waarden): heel sterk bivariaat verband

Er is in het voorbeeld tussen het plegen van delicten en het slachtoffer een matige positieve (lineaire) samenhang vast te stellen met een waarde van 0.44.

3.2. Asymmetrische associatiematen: bivariate regressie analyse

3.2.1. Te onthouden kernbegrippen

Afhankelijke variabele	Responsvariabele of explanandum
Determinatie coëfficiënt (R^2)	Een goodness of fit maat die weergeeft hoeveel procent van de geobserveerde verschillen of varia(n)tie in de afhankelijke variabele kan verklaard worden op basis van de onafhankelijke variabele. = proportie van de totale variatie in Y die door X kan verklaard worden. In de bivariate regressieanalyse is R^2 gelijk aan het kwadraat van de correlatiecoëfficiënt tussen X en Y.
Gestandaardiseerde regressiecoëfficiënt (β)	Geeft de sterke weer van een effect. Is in de bivariate regressie analyse gelijk aan de correlatiecoëfficiënt. Berekening: covariantie tussen X en Y delen door het product van standaardafwijking van X met standaardafwijking van Y. Neemt een waarde aan tussen -1 en +1 waarbij -1 en +1 een perfecte relatie aanduiden (dwz de onafhankelijk variabele kan de afhankelijke perfect voorspellen, dus geen residuen). Bij een waarde 0 heeft de onafhankelijke geen effect op de afhankelijke variabele.
Foutenterm of error	Foutenterm of residu of residuele term is het verschil tussen de werkelijk (geobserveerde) waarde van de afhankelijke variabele en de voorgespelde waarde van de afhankelijke variabele.
Intercept	Zie: regressieconstante
Onafhankelijke variabele	Predictor variabele of explanans

Ongestandaardiseerde regressiecoëfficiënt (b)	Geeft aan met hoeveel eenheden Y toeneemt als x met één eenheid toeneemt. Berekening: covariantie tussen X en Y gedeeld door de variantie in X.
Regressie analyse	Het, met een zekere mate van precisie, voorspellen van de score op een afhankelijke variabele (Y) op basis van één (bivariate regressie) of meerdere onafhankelijke (multiple regressie) X-variabelen.
Regressie coëfficiënten	Regressieconstante of het intercept en het regressiegewicht dat de hellingshoek van de regressielijn aanduidt.
Regressieconstante of het intercept (a)	De verwachte of voorspelde waarde van de afhankelijke variabele Y wanneer de waarde op de onafhankelijke variabele 0 bedraagt.
Regression sum of squares	Geeft de variatie in de afhankelijke variabele weer die voorspeld kan worden op basis van de onafhankelijke variabele. Wordt berekend door de afwijkingen van de voorspelde waarden van de afhankelijke variabele ten opzichte van het rekenkundig gemiddelde van de afhankelijke variabele te kwadrateren en vervolgens deze kwadraten te sommeren.
Residual sum of squares	De residuele variatie of residual sum of squares wordt berekend door de afwijkingen van de werkelijke waarde van de afhankelijke variabele ten opzichte van de voorspelde waarde van de afhankelijke variabele te kwadrateren en op te tellen.

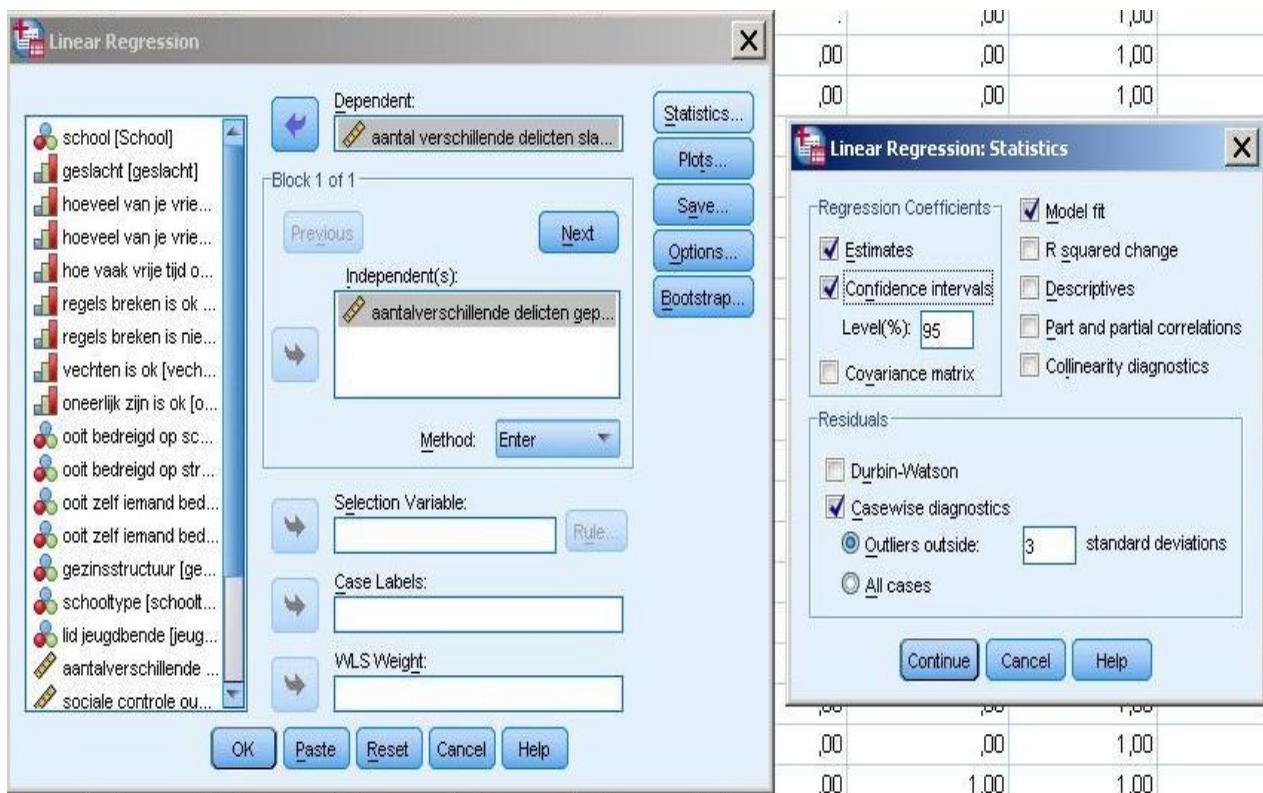
3.2.2. Statistische symbolen en formules

Determinatie coëfficiënt (R^2)	$R^2 = \frac{\sum_{i=1}^n [\hat{y}_i - \bar{Y}]^2}{\sum_{i=1}^n [y_i - \bar{Y}]^2}$
Enkelvoudige lineaire regressievergelijking	$Y = a + b_1 X + e$
Ongestandaardiseerde regressiecoëfficiënt (b)	$b_1 = r(X,Y) S_Y/S_X$
Regressieconstante of intercept (a)	$a = \bar{Y} - b_1 \bar{X}$

3.2.3. Asymmetrische associatiematen op metrisch niveau: SPSS-output en interpretatie (bivariate regressie analyse metrische variabelen)

Wie kwaad doet, kwaad ontmoet ?

Een zeer belangrijke tool in de criminologie is de lineaire regressieanalyse waar de onderzoeker geïnteresseerd is in het voorspellen van de waarden op de afhankelijke variabele op basis van een onafhankelijke variabele. We hanteren hier een populair vraagstuk uit de criminologie: er wordt wel eens beweerd, en we hebben het hierboven reeds aangetoond, dat er een verband bestaat tussen enerzijds het plegen van delicten en anderzijds het slachtoffer worden van delicten. Bij de berekening van de correlatiecoëfficiënt hebben we vastgesteld dat de samenhang in onze dataset inderdaad positief is: 0.44. De bivariate regressieanalyse is een asymmetrische analysetechniek die ons toelaat voorspellingen te maken voor de waarden op de Y-variabele (de afhankelijke variabele) op basis van de onafhankelijke predictor X. We moeten hier dus een expliciete keuze maken vanuit theoretische gronden. Welke variabele beschouwen we als afhankelijke en welke als onafhankelijke? We gaan er hier theoretisch van uit dat we het plegen van delicten ("kwaad doen") zien als onafhankelijke variabele. Wie kwaad doet, zou dan meer slachtoffer kunnen worden. Dit is een eenzijdige oorzakelijke interpretatie van een gekend verband uit criminologisch onderzoek, maar om didactische redenen houden we ons aan deze interpretatie. In SPSS voeren we een regressie-analyse uit door onder "analyze" te kiezen voor "linear regression".



De afhankelijke variabele brengen we via de overbrengknop onder “Dependent” en de onafhankelijke variabele brengen we via de overbrengknop onder “Independent”. Door op “options” te klikken, kunnen we de regressieparameters opvragen. We vragen systematisch de “estimates” (regressiecoëfficiënten), de betrouwbaarheidsintervallen², de “model fit” (determinatiecoëfficiënt), de beschrijvende statistieken en de residuele termen op.

De output van een regressie-analyse ziet er in SPSS als volgt uit:

Variables Entered/Removed ^b			
Model	Variables Entered	Variables Removed	Method
1	aantalverschillende delicten gepleegd ^a	.	Enter

- a. All requested variables entered.
- b. Dependent Variable: aantal verschillende delicten slachtoffer

² Hierop komen we verder terug.

SPSS informeert onder “Variables entered” welke de afhankelijke en onafhankelijke variabele is. Dit is belangrijk. Zo zien we of we geen vergissingen gemaakt hebben.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,444 ^a	,197	,196	1,84028

- a. Predictors: (Constant), aantal verschillende delicten gepleegd
- b. Dependent Variable: aantal verschillende delicten slachtoffer

Onder “model summary” staan achtereenvolgens de correlatiecoëfficiënt, de determinatiecoëfficiënt, de aangepaste determinatiecoëfficiënt en de standaardfout van de schatter. Deze twee laatste waarden worden later behandeld. De Anova tabel is zeer belangrijk bij de inferentiële statistiek en dit vormt een onderdeel van een volgend hoofdstuk. Een aantal zaken zullen we hier reeds bespreken.

Uit de “*model summary*” of de samenvatting van de fit van het statistische model, zien we de volgende coëfficiënten:

R: deze is de samenhang tussen de geobserveerde waarde en de voorspelde waarde op de Y-variabele. Deze waarde is hier gelijk aan de bivariate correlatiecoëfficiënt tussen X en Y.

R Square: deze is de determinatiecoëfficiënt. Het is de verklaarde variantie: 19.7% van de geobserveerde verschillen op de afhankelijke variabele (slachtofferschap) kan verklaard worden vanuit het plegen van delicten.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1203,779	1	1203,779	355,451	,000 ^a
	Residual	4913,997	1451	3,387		
	Total	6117,776	1452			

- a. Predictors: (Constant), aantal verschillende delicten gepleegd
- b. Dependent Variable: aantal verschillende delicten slachtoffer

Uit de ANOVA tabel kunnen we hetzelfde aflezen als wat erboven stond:

We krijgen zicht op de “Regression sum of squares”, de “Residual sum of squares” en de “Total sum of squares”.

De formule van de determinatiecoëfficiënt zegt:

$$R \text{ square is } \frac{\sum_{i=1}^n [\hat{y}_i - \bar{Y}]^2}{\sum_{i=1}^n [y_i - \bar{Y}]^2}$$

Ofwel: "Neem de som van het kwadraat van het verschil tussen elke voorspelde waarde van Y op basis van X (\hat{Y} of Y -hat) minus het gemiddelde op Y (= "regression sum of square")."

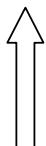
"En deel deze som vervolgens door de som van het kwadraat van het verschil tussen alle geobserveerde waarden van Y minus het gemiddelde op Y (= "total sum of squares" of de variatie in Y)."

Dus: $1203.779/6117.776 = 0.197$ of 19.7% van de waargenomen verschillen in slachtofferschap kan verklaard worden door de waargenomen verschillen in delicten plegen.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1 (Constant)	1,449	,061		23,825	,000	1,330	1,568
aantalverschillende delicten gepleegd	,475	,025	,444	18,853	,000	,426	,525

a. Dependent Variable: aantal verschillende delicten slachtoffer



schattingen intercept en richtingscoef.

geschatte regressielijn is =

$\hat{Y} = 1.449 + 0.475 X$

De belangrijkste informatie staat in de output onder "coefficients". Hier lezen we de regressiecoëfficiënten af. We zien dat SPSS standaard het intercept en de hellingshoek (richtingscoëfficiënt) weergeeft. SPSS geeft zowel de ongestandaardiseerde coëfficiënten als de gestandaardiseerde coëfficiënten.

De constante (a of b_0) of het intercept bedraagt 1.449. Dit is het verwacht aantal keer dat iemand slachtoffer wordt voor iemand die geen enkele keer een crimineel feit heeft gepleegd.

De ongestandaardiseerde richtingscoëfficiënt (b_1) bedraagt 0.475. SPSS geeft deze weer met de benaming B. Dit is de verwachte toename in Y als X met een eenheid stijgt, dus:

$$\hat{Y} = b_0 + b_1 X$$

Of: $\hat{Y} = 1.449 + 0.475$ (aantal verschillende delicten gepleegd). We kunnen dus door een waarde op X (aantal delicten gepleegd) in te vullen, te weten komen hoeveel keer iemand verwacht wordt slachtoffer te worden. Voor iemand die 3 delicten heeft gepleegd, wordt deze verwachting: $1.449 + 0.475 * 3$.

De gestandaardiseerde richtingscoëfficiënt komt in een bivariate analyse overeen met de bivariate correlatiecoëfficiënt. De overige parameters worden in een later hoofdstuk beschreven.

Interessant is dat SPSS ons de mogelijkheid biedt om “outliers” of “uitbijters” te bekijken. Dit zijn observaties waar we een erg slechte voorspelling hebben gedaan. Met erg slecht bedoelen we dat we er met de voorspelling zeer ver naast zaten: minstens drie standaardafwijkingen (zie ook het hoofdstuk over de normale verdeling en standaardnormale scores). Dit is interessant omdat we deze cases afzonderlijk in de diepte kunnen bestuderen.

Casewise Diagnostics^a

Case Number	Std. Residual	aantal verschillende delicten slachtoffer	Predicted Value	Residual
65	4,130	10,00	2,3998	7,60020
154	3,301	8,00	1,9245	6,07554
159	3,560	8,00	1,4491	6,55088
210	4,647	10,00	1,4491	8,55088
216	3,043	8,00	2,3998	5,60020
295	3,016	7,00	1,4491	5,55088
358	3,845	9,00	1,9245	7,07554
431	3,043	8,00	2,3998	5,60020
640	4,388	10,00	1,9245	8,07554
700	4,130	10,00	2,3998	7,60020
926	-3,370	,00	6,2025	-6,20252
1142	3,016	7,00	1,4491	5,55088
1145	3,560	8,00	1,4491	6,55088
1351	3,613	10,00	3,3505	6,64952
1379	-3,112	,00	5,7272	-5,72718
1523	3,016	7,00	1,4491	5,55088
1553	3,355	10,00	3,8258	6,17418

a. Dependent Variable: aantal verschillende delicten slachtoffer

Residuals Statistics^a

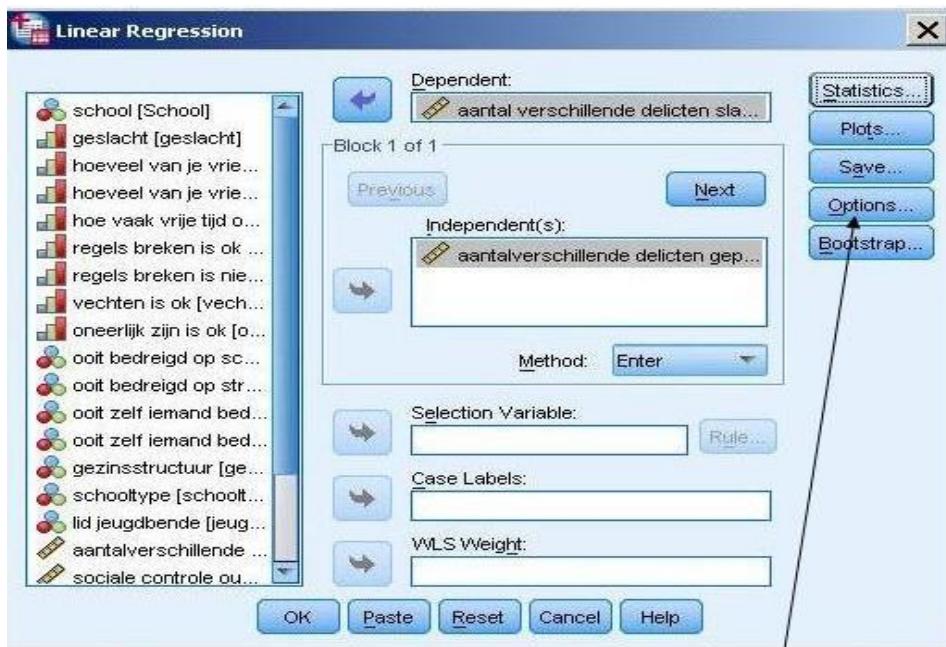
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,4491	7,6285	2,1466	,91052	1453
Residual	-6,20252	8,55088	,00000	1,83965	1453
Std. Predicted Value	-,766	6,021	,000	1,000	1453
Std. Residual	-3,370	4,647	,000	1,000	1453

a. Dependent Variable: aantal verschillende delicten slachtoffer

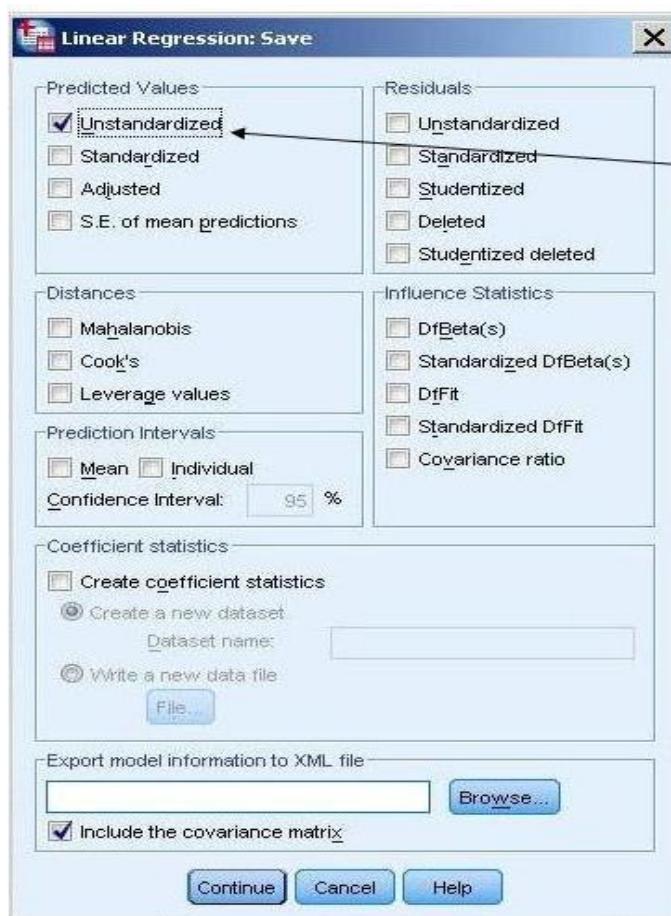
Tot slot geeft SPSS ook de residuele statistieken (minimum, maximum, gemiddelde, standaardafwijking). We zien achtereenvolgens de voorspelde waarde (predicted value), de residuele, de gestandaardiseerde voorspelde waarde (deze heeft natuurlijk een gemiddelde van nul en std van 1), precies zoals de gestandaardiseerde residuele term. De “residual statistics” zijn de beschrijvende statistieken van de residuele termen. Een residuele term is het verschil tussen wat er geobserveerd werd als waarde op Y en wat we verwachtten als waarde op Y op basis van onze kennis over X.

Zelf stapsgewijs narekenen van de determinatiecoëfficiënt aan de hand van SPSS

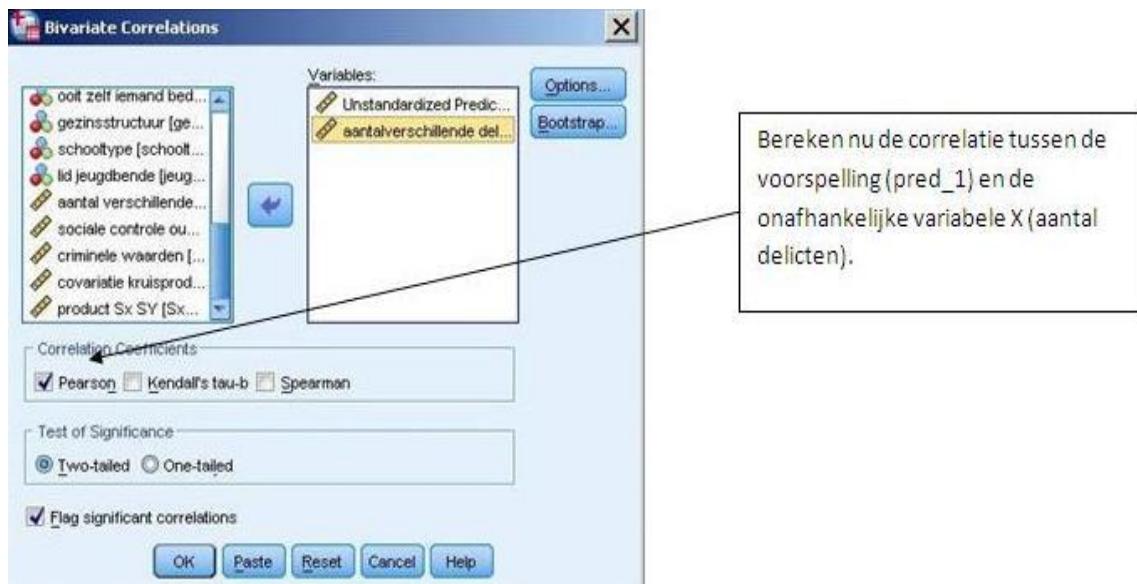
De determinatiecoëfficiënt is ook het kwadraat van correlatie tussen de geobserveerde waarden op Y en de voorspelde (of verwachte) waarde voor Y op basis van X. Laten we dit even zelf narekenen. We tonen aan dat SPSS de formule inderdaad correct toepast. Zo gaan we te werk:



Klik op "options" en bewaar de verwachte waarde van Y op basis van X



Opgelet! Deze voorspellingen op basis van X worden door SPSS weggeschreven (of bewaard) onder de naam Pre_1. Je kan zelf deze naam veranderen als je wil.



Lees het resultaat af uit de tabel.

Correlations

		Unstandardized Predicted Value	aantal verschillende delicten slachtoffer
Unstandardized Predicted Value	Pearson Correlation	1	,444**
	Sig. (2-tailed)		,000
	N	1504	1453
aantal verschillende delicten slachtoffer	Pearson Correlation	,444**	1
	Sig. (2-tailed)	,000	
	N	1453	1491

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

Hiermee hebben we aangetoond dat R inderdaad de samenhang is tussen de voorspelling en de geobserveerde waarde. Nemen we daarvan het kwadraat, dan bekomen we inderdaad de determinatiecoëfficiënt.

4. SAMENVATTENDE TABEL

NOMINAAL MEETNIVEAU	ORDINAAL MEETNIVEAU	METRISCH MEETNIVEAU
Percentageverschil (uitgedrukt in percentagepunten)	Gamma -1 tot +1	Covariatie
Odds (ratio) 1= geen associatie Van 0 tot $+\infty$	Spearmans rho -1 tot +1	Covariantie
Chi ² Van 0 tot $+\infty$	Kendall's Tau-b -1 tot +1	Correlatie-coëfficiënt
Phi (2*2 tabel) 0 tot 1		
Cramer's V (r*k tabel) 0 tot 1		

5. OEFENINGEN

- De Chi²-toets is één van de meest gebruikte manieren om relaties tussen twee of meer categorische variabelen te bestuderen.

Onderzoekers verzamelden gegevens over rookstatus en de diagnose longkanker bij een willekeurige steekproef van volwassenen. Elk van deze variabelen is dichotoom: een persoon rookt momenteel of niet en heeft een longkankerdiaagnose of niet.

<i>Longkanker diagnose</i>			
Rookstatus	Diagnose	Geen diagnose	Totaal
Roker	60	300	
Niet-roker	10	390	
			<i>N=</i>

- Hoeveel bedraagt de marginale frequentieverdeling voor de variabele longkanker diagnose?
- Hoeveel bedraagt de marginale frequentieverdeling voor de variabele Rookstatus?
- Hoeveel procent van de respondenten met longkanker diagnose is roker?
- Hoeveel procent van de respondenten zonder longkanker diagnose is roker?
- Hoeveel bedraagt het relevante percentageverschil?

Bereken Chi²

- Bereken de verwachte waarden/frequenties voor elke cel (dit zijn de verwachte frequenties wanneer de twee variabelen onafhankelijk zijn of niet samenhangen)

<i>Longkanker diagnose</i>			
Rookstatus	Diagnose	Geen diagnose	Totaal
Roker	60	300	
Niet-roker	10	390	
			<i>N=</i>

- Hoeveel bedraagt Chi²?
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen?
- Hoeveel bedraagt deze? Interpreteer.

2. Is er een verband tussen vuurwapenbezit en biologisch geslacht?

In een hypothetisch onderzoek wordt aan 817 mannen en aan 1040 vrouwen gevraagd of zij een vuurwapen bezitten. De onderzoeker wil weten of er een verband bestaat tussen het bezit van een vuurwapen en biologisch geslacht. Hieronder vind je de frequenties. Maak een volledige analyse van de kruistabel.

Opgelet! Bepaal vooraf welke variabelen je zou beschouwen als onafhankelijk en als afhankelijk. Plaats de onafhankelijke in de kolommen (=kolomvariabele) en de afhankelijke in de rijen (rijvariabele).

<i>Ben jij in het bezit van een vuurwapen?</i>		
Vuurwapenbezit	JA	NEEN
Man	343	474
Vrouw	260	780

- Bereken het relevante percentageverschil.
- Bepaal de H_0 en H_a .
- Bereken χ^2 .
- Welke associatiemaat gebaseerd op χ^2 kun je bijkomend berekenen?
- Hoeveel bedraagt deze? Interpreteer.

3. *Is er een verband tussen het roken van sigaretten en het drinken van alcohol bij studenten?*

Hieronder vind je een 2*2 kruistabel. Aan 110 studenten werd de vraag gesteld of ze al dan niet rookten en of ze al dan niet dronken.

		Rook je?	
		JA	NEE
Drink je?	JA	50	15
	NEE	20	25

- Bereken het relevante percentageverschil.
- Bepaal de H₀ en H_a.
- Bereken Chi².
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen ?
- Hoeveel bedraagt deze ? Interpreteer.

4. Is er een verband tussen burgerlijke staat en drinkgedrag?

Een nationale enquête werd uitgevoerd om informatie te verkrijgen over de alcoholconsumptiepatronen van volwassenen in Vlaanderen op basis van hun burgerlijke staat. Een willekeurige steekproef van 1772 inwoners van 18 jaar en ouder leverde de weergegeven gegevens op:

		Aantal glazen alcohol per maand			
		0	1-60	Meer dan 60	Totaal
Burgerlijke staat	Vrijgezel	67	213	74	354
	Gehuwd	411	633	127	1173
	Weduwe weduwnaar	85	51	7	143
	Gescheiden	27	60	15	102
Totaal		590	957	225	1772

- Bepaal de H0 en Ha.
- Bereken Chi².
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen ?
- Hoeveel bedraagt deze ? Interpreteer.

5. *Angst voor jeugdbendes.*

De relatie tussen *bezorgdheid voor diefstal door jeugdbendes* en *daadwerkelijk slachtoffer worden van diefstal* werd onderzocht. Interviews met een steekproef van middelbare scholieren leverden de volgende kruistabel op:

		<i>Bezorgdheid voor diefstal door jeugdbendes</i>		TOTAAL
		JA	NEE	
<i>Angst om daadwerkelijk slachtoffer te worden</i>	JA	58	45	
	NEE	29	71	
	TOTAAL			

- Bereken Chi².
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen ?
- Hoeveel bedraagt deze ? Interpreteer.

6. *Is er een verband tussen ‘ontrouw’ en ‘geluk in de relatie’?*

Onderstaande kruistabel geeft data over het aantal mannen dat al dan niet ontrouw is, in relatie tot de mate waarin zij aangeven gelukkig te zijn in hun relatie. De variabele ‘geluk in de relatie’ werd hercodeerd naar een dichotome variabele op basis van de mediaan. Scores hoger dan de mediaan werden gecodeerd als ‘ongelukkig’ en scores gelijk aan of lager dan de mediaan werden gecodeerd als ‘gelukkig’.

		<i>Ontrouw?</i>		TOTAAL
		Ontrouw	Trouw	
<i>Geluk in de relatie</i>	Ongelukkig	56	101	157
	Gelukkig	62	287	349
	TOTAAL	118	388	506

- Hoeveel procent van het aantal mannen die rapporteren gelukkig te zijn in hun relatie, is ontrouw?
- Hoeveel procent van het aantal mannen die rapporteren trouw te zijn in hun relatie, is ongelukkig?
- Is er een verband tussen ontrouw en geluk in de relatie?
- Hoe sterk is het verband?

7. De Vlaamse minister van onderwijs start een grootschalig onderzoek op naar pestgedrag in het Vlaamse secundair onderwijs. Check-It, het ondertussen wereldvermaarde onderzoeksbureau, bezorgt binnen de kortste keren het onderzoeksrapport aan de minister. Eén van de resultaten belicht het eventuele geslachtsverschil in het pesten.

	Jongens	Meisjes	Totaal
Niet-pester	4613	5530	10143
Pester	1132	568	1700
Totaal	5745	6098	11843

- Bereken de oddsratio van jongens tegenover meisjes (pesters vs niet-pesters).

- Wat zegt de zojuist berekende oddsratio van jongens tegenover meisjes (pester vs. Niet-pesters) ?
 - a. Er zijn in totaal 2,39 maal minder pesters dan niet-pesters (zowel onder de meisjes als onder de jongens)
 - b. De verhouding tussen pesters en niet-pesters ligt 2,39 keer hoger bij de jongens dan bij de meisjes
 - c. De verhouding tussen pesters en niet-pesters ligt 2,39 keer hoger bij de meisjes dan bij de jongens
 - d. Er zijn in totaal 2,39 maal meer pesters dan niet-pesters (zowel onder de meisjes als onder de jongens)

8. Beantwoord onderstaande vragen

	Jongens	Meisjes	Totaal
Ooit alcohol gebruikt	410	472	
Nooit alcohol gebruikt	623	1263	
Totaal			

- a. Bepaal voor mannen de odds op 'ooit gebruikt' t.o.v. 'nooit gebruikt' en interpreteer

- b. Bepaal voor vrouwen de odds op 'ooit gebruikt' t.o.v. 'nooit gebruikt' en interpreteer

- c. Bepaal vervolgens de oddsratio mannen t.o.v. vrouwen. Druk de betekenis uit in je eigen woorden.

- d. Bepaal vervolgens de oddsratio vrouwen t.o.v. mannen. Druk de betekenis uit in je eigen woorden

9. Lees onderstaande uitspraken. Elk van de uitspraken bevat een blunder. Geef voor elk geval aan wat er mis is.

'Er bestaat een grote correlatie tussen het geslacht van Amerikaanse werknemers en hun inkomen.'

'We vonden een grote correlatie ($r = 1.09$) tussen de door studenten gegeven beoordeling over de kwaliteit van het lesgeven van stafleden en de door andere stafleden gegeven beoordeling'.

'De correlatie tussen plantdichtheid en de opbrengst van maïs bleek $r = 0.23$ kubieke meter te zijn'.

10. Onze emotionele reactie op sociale afwijzing wordt omschreven als 'sociale pijn'. Een onderzoek bekeek of sociale afwijzing zorgt voor activiteit in hersengebieden waarvan bekend is dat deze door fysieke pijn worden geactiveerd. Als dat het geval is, dan ervaren we werkelijk sociale en fysieke pijn op dezelfde manier. Personen in het onderzoek werden eerst bij een sociale activiteit betrokken en vervolgens opzettelijk buitengesloten terwijl de toename van de bloedcirculatie in hun hersenen werd gemeten. Na elke activiteit vulden de personen een vragenlijst in om aan te geven in welke mate zij zich buitengesloten voelden.

Hieronder volgen de gegevens voor 13 proefpersonen. De verklarende variabele is 'sociaal leed'. Deze variabele is gemeten aan de hand van de scores in elke vragenlijst, waarbij de score na de buitensluiting wordt gerelateerd aan de score na afloop van de activiteit. Dus, waarden groter dan 1 tonen de mate van pijn veroorzaakt door het buitensluiten. De te verklaren variabele is de activiteit in de anterior cingulate cortex, een hersengebied dat door fysieke pijn wordt geactiveerd.

Proefpersoon	Sociaal leed	Hersenactiviteit
1	1.26	-0.055
2	1.85	-0.040
3	1.10	-0.026
4	2.50	-0.017
5	2.17	-0.017
6	2.67	0.017
7	2.01	0.021
8	2.18	0.025
9	2.58	0.027
10	2.75	0.033
11	2.75	0.064
12	3.33	0.077
13	3.65	0.124

OPGAVE

Beschrijf de richting en de sterkte van de relatie tussen sociaal leed en hersenactiviteit. Suggereren de gegevens dat hersenactiviteit in het 'pijn'-gebied werkelijk direct is gerelateerd aan pijn vanwege sociale uitsluiting?

11. De stofwisselingssnelheid, de snelheid waarmee het lichaam energie verbruikt, is van belang bij onderzoek naar gewichtstoename, diëten en lichaamsbeweging. In onderstaande tabel vind je de gegevens over het vettvrij lichaamsgewicht (=gewicht van een persoon zonder vet) en de stofwisselingssnelheid in rust voor 7 vrouwen en 7 mannen die als proefpersoon bij een onderzoek naar afvallen zijn betrokken. De stofwisselingssnelheid wordt gemeten als het aantal calorieën dat per 24 uur wordt verbrand. De onderzoekers denken dat het vettvrij lichaamsgewicht een belangrijke invloed op de stofwisselingssnelheid heeft.

Proefpersoon	Biologisch geslacht	Vetvrij lichaamsgewicht	Stofwisselingssnelheid
1	M	62.0	1792
2	M	62.9	1666
3	V	36.1	995
4	V	54.6	1425
5	V	48.5	1396
6	V	42.0	1418
7	M	47.4	1362
8	V	50.6	1502
9	V	42.0	1256
10	M	48.7	1614
11	V	40.3	1189
12	M	51.9	1460
13	M	51.9	1867
14	M	46.9	1439

OPGAVE:

- Is de samenhang tussen ‘vetvrij lichaamsgewicht’ en ‘stofwisselingssnelheid’ positief of negatief?
- Hoe sterk is de samenhang?
- Is de richting en de sterkte van de samenhang verschillend voor mannen en vrouwen?
- Bepaal de helling van de regressielijn voor de stofwisselingssnelheid ten opzicht van het vettvrij lichaamsgewicht.
- Bepaal de helling van de regressielijn voor het vettvrije lichaamsgewicht ten opzichte van de stofwisselingsratio.

12. Voor verschillende ontwikkelingslanden is koffie een belangrijk exportartikel. Wanneer de koffieprijzen hoog zijn, kappen de boeren vaak bossen om meer koffiebomen te planten. Hieronder vind je de gegevens van de prijzen die aan koffieverbouwers in Indonesië werden betaald en de mate van ontbossing in een nationaal park dat in een koffieproducerende regio ligt, beide over een periode van vijf jaar.

PRIJS (dollarcenten per Am. pond ³)	ONTBOSSING (percent)
29	0.49
40	1.59
54	1.69
55	1.82
72	3.10

OPGAVE

- Wat is de verklarende variabele?
- Hoeveel bedraagt Pearson's correlatie?
- De prijs van koffie werd uitgedrukt in dollars. Als de koffie in euro's was geprijsd en de dollarprijzen in de tabel hierboven waren vertaald in equivalenten euro's, zou de correlatie tussen koffieprijs en percentage ontbossing dan veranderen? Verklaar jouw antwoord.

³ 1 Amerikaanse pond is 0.45359237 kg.

13. Een studente vraagt zich af of mensen van overeenkomstige lengte geneigd zijn met elkaar uit te gaan. Ze meet haar eigen lengte, die van haar kamergenote en die van de vrouwen in de naastgelegen kamers. Vervolgens meet ze de lengte van de eerste man waarmee elke vrouw uitgaat. Hieronder zijn de data (lengte in inches⁴):

Vrouwen	66	64	66	65	70	65
Mannen	72	68	70	68	71	65

OPGAVE

- Bepaal de correlatie tussen de lengtes van mannen en vrouwen.
- Als de lengtes in centimeters waren gemeten, in plaats van in inches, in welk opzicht verandert de correlatie dan?

⁴ 1 inch is 2.54 cm.

14. WAAR of FOUT ?

Uitspraak	WAAR	FOUT
Als we de samenhang berekenen tussen twee nominale variabelen met elk twee categorieën, dan bedraagt de waarde voor Cramers V dezelfde als Phi.		
Cramers V is gebaseerd op Chi-kwadraat.		
Chi-kwadraat is een goodness-of-fit maat.		
De regressiecoëfficiënt bepaalt of je rechte daalt of stijgt. b > 0 : rechte stijgt b < 0 : rechte daalt		
Als $R^2 = 1$: is Y perfect onvoorspelbaar.		
Bij variabelen gemeten op twee verschillende parametrische meetniveaus dient men sensu stricto nominale associatiematen te gebruiken.		
De gekwadrateerde correlatiecoëfficiënt R^2 zegt welk deel van de variantie in Y verklaard wordt door X.		
Als variabele X gemeten is op het nominale niveau en variabele Y op het ordinale niveau, dan is gamma de beste oplossing.		
Spearman's rho is gebaseerd op de Pearson's r, maar dan voor ordinale variabelen.		
Bij een r^* k tabel is Phi gelijk aan V.		
Het is zinloos om de product-moment correlatiecoëfficiënt r te berekenen als je een curvilineair verband ontdekt.		
De verhouding tussen de regression sum of squares en de residual sum of squares is gelijk aan de determinatiecoëfficiënt.		
De residuale term is het verschil tussen de observatie en de voorspelling.		
Positieve residuen liggen boven de regressielijn.		

15. Blijf je slank door wiebelen en draaien?

Sommige mensen worden niet zwaarder, zelfs als zij zich overeten. Een mogelijke verklaring is dat activiteiten als wiebelen en draaien en andere niet-sportmatige activiteiten hieraan ten grondslag liggen.

Onderzoekers gaven 16 gezonde jonge volwassenen 8 weken lang te veel voeding. Zij maten de *gewichtstoename* (in kilo's) en hanteerden als een verklarende variabele de toename in *energieverbruik* (in calorieën) uit activiteiten zoals wiebelen en draaien en andere niet-sportmatige dagelijkse activiteiten. Hieronder vind je de gegevens

ENERGIEVERBRUIK (in calorieën)	GEWICHTSTOENAME (in kilo's)
-94	4.2
-57	3.0
-29	3.7
135	2.7
143	3.2
151	3.6
245	2.4
355	1.3
392	3.8
473	1.7
486	1.6
535	2.2
571	1.0
580	0.4
620	2.3
690	1.1

OPGAVE

- Wat is de onafhankelijke variabele?
- Wat is de afhankelijke variabele?
- Hoeveel bedraagt Pearson's correlatiecoëfficiënt? Interpreteer.
- Bepaal de bivariate regressievergelijking.
- Wat is de meeteenheid van de ongestandaardiseerde richtingscoëfficiënt.
- Interpreteer de betekenis van de ongestandaardiseerde richtingscoëfficiënt.
- Hoeveel bedraagt de gewichtstoename voor een individu van wie het energieverbruik met 400 calorieën toeneemt?

16. Een universiteitskrant interviewt een psycholoog over de door studenten beoordeelde kwaliteit van het lesgeven door stafleden. De psycholoog zegt: "*De gegevens tonen aan dat de correlatie tussen de onderzoeksproductiviteit en de onderwijskundige beoordeling van een staflid dicht bij 0 ligt.*" De krant maakt hier het volgende van: "*Professor McDaniel vindt dat goede onderzoekers meestal slechte leraren zijn, en andersom.*"

Leg uit waarom dit bericht onjuist is. Geef in jouw eigen woorden de bedoelingen van de psycholoog weer.

17. Controlevragen

- Waar staan in een kruistabel de afhankelijke en de onafhankelijke variabele meestal?
- Stel, je brengt de gegevens van de vakantiekeuze van 1^{ste} Bac-Criminologie studenten in kaart (goedkoop versus duur). Ook weet je of deze studenten een studielening hebben of niet. Waar zet je deze variabelen in de kruistabel?
- Stel dat je wilt weten of er onder de studenten die een lening hebben, vaker dure dan goedkope vakantie worden geboekt. Welke manier van percenteren kies je dan (rijpercentage of kolompercentages)?
- Wanneer kun je een Chi-kwadraattoets inzetten?
- Welk meetniveau moeten de variabelen minimaal hebben?
- Stel, je wilt de samenhang onderzoeken tussen aantal jaren opleiding en inkomen. Kun je dan ook een Chi-kwadraattoets gebruiken?
- Hoe interpreteer je de Spearman-rangcorrelatie in vergelijking met de Pearson correlatiecoëfficiënt?
- Stel, je vindt een correlatiecoëfficiënt van 0.45. Hoe sterk is het gevonden verband ?
- Welke waarden kan een correlatiecoëfficiënt aannemen?

- Stel, je vindt tussen ‘fietsen’ en ‘conditie’ een verband van 0.70. Hoe zou je dit interpreteren?
- Wat is het verschil tussen ‘een verband’ en ‘een effect’?
- Wanneer gebruik je een regressieanalyse?
- Wat betekent ‘de constante’ in de regressievergelijking?
- Wat is de functie van een regressiecoëfficiënt?
- Hoe kun je de verklaarde variantie definiëren?
- Wat zegt de verklaarde variantie over het regressiemodel?
- Stel, je vindt een R^2 van 60%. Hoe sterk is je model dan?
- Hoe reken je de residuele waarde uit?
- Stel dat je het *relatieve netto*-effect van een variabele wilt analyseren in relatie tot andere effecten. Welke coëfficiënt gebruik je dan?
- Wat is het verschil in interpretatie tussen de B en Beta-coëfficiënt?

18. UITSPRAKEN: JUIST OF FOUT?

<i>Hangt het dagelijks eten van fruit samen met een betere gezondheid (gemeten aan de hand van aantal ziektedagen)?</i>	WAAR	FOUT
EFFECT VAN FRUIT OP ZIEKTE		
	Beta	b
FRUIT	-0,58	-1,14
<i>Afgerond 34% van de variantie in ziekte kan verklaard worden door de variantie in het eten van fruit.</i>		
<i>Voor iemand die 0 dagen thuisblijft wegens ziekte, verwachten we dat hij/zij gemiddeld 6.21 stukken fruit eet.</i>		
<i>Bij toename van 1 standaardafwijking in het aantal stukken fruit dat men eet, verwachten we een afname met 1.14 standaardafwijkingen in aantal ziektedagen.</i>		
<i>Joris at gemiddeld 2.5 stukken fruit per dag en bleef 4 dagen thuis. Het residu bedraagt afgerond 0.64.</i>		
<i>Als Els gemiddeld per dag 2 stukken fruit meer eet dan Marie, dan verwachten we dat Els afgerond 2 dagen minder thuis blijft wegens ziekte dan Marie.</i>		

HOOFDSTUK 6

INFERENTIELLE STATISTIEK

HYPOTHESIS TESTING

NOW WE ENTER A NEW AREA... GOVERNMENT, BUSINESS, AND THE HARD AND SOFT SCIENCES ALL USE AND OFTEN ABUSE THESE TESTS OF SIGNIFICANCE. IT'S ALL ABOUT ANSWERING THE QUESTION, "COULD THESE OBSERVATIONS REALLY HAVE OCCURRED BY CHANCE?"



HOOFDSTUK VI

INFERENTIELE STATISTIEK

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk kunnen studenten de centrale begrippen die verband houden met de inferentiële statistiek zoals schatten, toetsen, p-waarden, nulhypothese, alternatieve hypothese, betrouwbaarheidsinterval toepassen. Het principe van de centrale limietstelling is zeer goed begrepen. Studenten kunnen betrouwbaarheidsintervallen met de hand uitrekenen en interpreteren. Studenten zijn in staat om vragen te beantwoorden aan de hand van SPSS-output. Studenten weten dat bij het toetsen niet enkel de normale verdeling wordt gebruikt maar ook andere verdelingen. Op basis van deze verdelingen kunnen we de kansen nagaan dat een bepaalde parameter in de populatie waaruit de steekproef afkomstig is effectief bestaat.

2. TE ONTHOUDEN KERNBEGRIPPEN

Alternatieve hypothese (H_a)	Hypothese waarvan we vermoeden dat ze waar is
Aselect	Toevallig, random. Een aselecte steekproef is een toevallig genomen steekproef
Betrouwbaarheidsinterval	Interval berekend uit steekproefdata volgens methode die bepaalde kans heeft een interval op te leveren waarin de populatiwaarde ligt Schatting +/- foutenmarge
Centrale Limietstelling	De eigenschap dat een steekproevenverdeling zo goed als normaal verdeeld is, ongeacht de vorm van de verdeling van de oorspronkelijke variabele.
Enkelvoudige aselecte steekproef (EAS)	Simple random sample
Inferentiële statistiek	Verzameling van statistische toetsen die via het toetsen van hypotheses en onderzoek van steekproeven, generaliserende uitspraken doen over populaties. Ook: verklarende, toetsende of analytische statistiek. Gaat dus een stap verder dan de beschrijvende statistiek die geen generaliseringen doet
Intervalschatting	Inschatten van de marges waarbinnen we met een zekere graad van onzekerheid een puntschatting inschatten.

Nulhypothese (H_0)	H_0 : 'geen verschil' of 'geen effect'
Overschrijdingskans	Kans, onder de aanname dat H_0 waar is, dat toetsingsgrootheid een waarde zou aannemen die gelijk is aan of nog groter of kleiner is (meer extreem is) dan de waargenomen waarde
Populatieparameter	Statistieken afkomstig uit de populatie (vb. gemiddelde, standaardafwijking)
Populatieverdeling	Verdeling van een variabele in de een populatie
Puntschatting	Schatting van een kenmerk in de populatie op basis van steekproefgegevens
Random sample	Toevalsstekproef waarin elke elementaire eenheid uit de empirische populatie een berekenbare kans heeft om in de steekproef te worden opgenomen
Representatief Schatten	Zeer goede afspiegeling De waarde van een steekproefkenmerk wordt gebruikt om een uitspraak te doen over een populatiekenmerk. Zo'n uitspraak is altijd gebaseerd op kansrekening.
Significant	Als een uitkomst in sterke mate de veronderstelling ondersteund dat het verschil niet door toeval is ontstaan
Significantie toets	Sterkte van bewijs tegen de nulhypothese (H_0)
Significantie niveau	α : significantieniveau: vaak 0,05 (of 0,01): bij onbeperkt aantal steekproeftrekkingen, verwerpen we in 5% van de gevallen H_0 foutief
Standaardfout (SE Standard Error)	Als de overschrijdingskans kleiner dan of gelijk is aan het significantie niveau, spreken we van statistische significantie ('bewijskracht voor het verwerpen van de H_0)
Steekproefgrootte	Standaardafwijking van het gemiddelde van een steekproevenverdeling
Steekproefverdeling	Verdeling van variabele in de steekproef
steekproevenverdeling (sampling distribution)	Geeft weer hoe steekproefgrooteden variëren bij onbeperkt aantal herhaalde steekproeftrekkingen (theoretische verdeling) uit zelfde populatie metzelfde n
Type I-fout	H_0 verwerpen terwijl ze juist is
Type II-fout	H_0 aanvaarden terwijl H_a juist is

3. STATISTISCHE SYMBOLEN EN FORMULES

Betrouwbaarheidsniveau :	$C = 1 - \alpha$ (α : kans op vergissing) $C = 90\% \rightarrow z^* = 1,645$ $C = 95\% \rightarrow z^* = 1,960$ $C = 99\% \rightarrow z^* = 2,576$
Foutenmarge	$m = z^* \frac{\sigma}{\sqrt{n}} \rightarrow n = \left(\frac{z^* \sigma}{m} \right)^2$
Opstellen betrouwbaarheidsinterval	$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$
Standaardfout	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Variantie van de steekproevenverdeling = populatievariantie delen door de steekproefomvang	$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

4. CENTRALE LIMIETSTELLING: DIDACTISCH VOORBEELD

In dit deel brengen we de abstracte theorie van de centrale limietstelling en de steekproeventheorie tot leven aan de hand van een didactisch voorbeeld. We doen dit aan de hand van populatiegegevens waaruit we zullen steekproeven leren trekken. De gegevens die we hebben verzameld zijn de slaagcijfers van studenten uit de eerste kandidatuur criminologische wetenschappen aan de Ugent tijdens het academiejaar 2001-2002. We hoeven dus geen steekproef te trekken, want de studentenadministratie verzamelt al deze gegevens. Om de theorie van de steekproevenverdeling uit te leggen, gaan we de redenering omdraaien. We gaan uit de gekende populatie verschillende toevalssteekproeven trekken, en de resultaten van deze steekproeven met elkaar leren vergelijken: we focussen in de voorbeelden vooral op het gemiddelde en de standaardafwijking die we noteren van elke toevalssteekproef en we kijken naar het gedrag van deze steekproefgrootheden: hoe ziet deze statistische verdeling of distributie van de beide kenmerken er uit? In de theorie hebben we gezegd dat steekproevenverdelingen voldoen aan bepaalde wetmatigheden waardoor je soms kan beredeneren hoe ze eruit zien: (1) uit de Centrale Limiet Stelling (CLS) weten we dat de steekproevenverdeling van het gemiddelde bij benadering een normaalverdeling is als

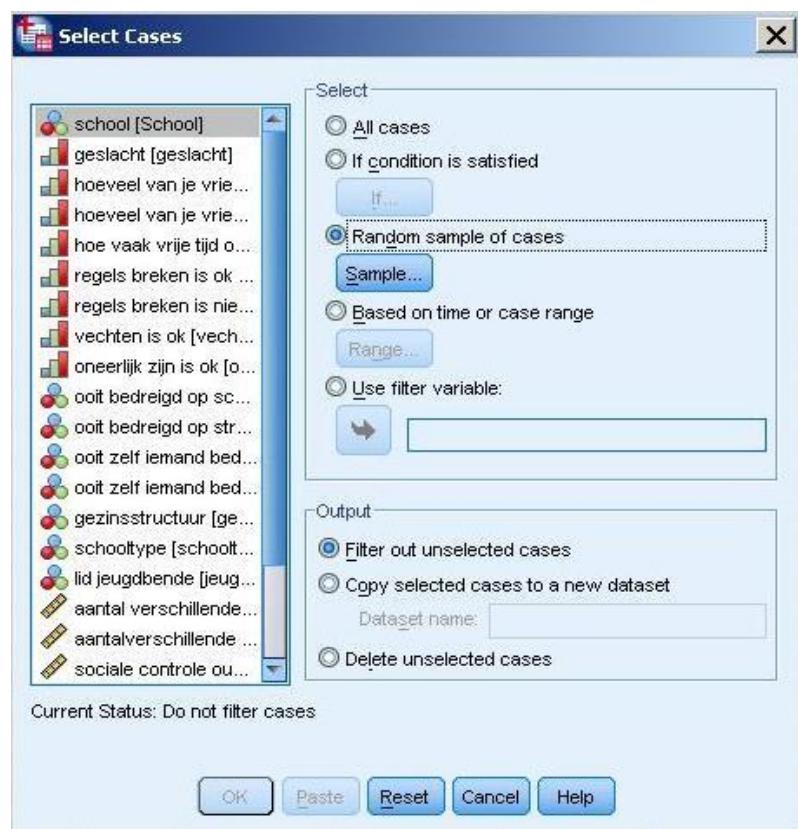
de steekproefomvang groot genoeg is. Volgens een vuistregel mag men ervan uitgaan dat de benadering goed is vanaf N=30. De wet van de grote getallen zegt dat grote steekproeven betere resultaten genereren. Tijd dus om de proef op de som te nemen. Het gemiddelde en de standaardafwijking kunnen worden afgelezen uit de onderstaande tabel:

Beschrijvende statistieken slaagcijfers op 1000 voor alle 1^{ste} jaarsstudenten criminologie

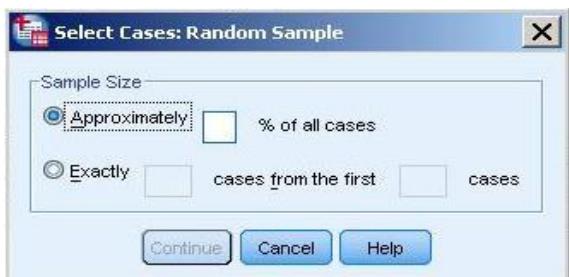
(Populatie= 1^{ste} kandidatuur criminologie Ugent)

	N Statistic	Mean Statistic	Std. Error	Deviation Statistic
Resultaat 1ste zittijd op 1000 punten	426	389,5869	10,87139	224,38296
Valid N (listwise)	426			

Om een toevalsssteekproef te trekken, kiezen we in SPSS onder “data” voor “select cases”.



Vink aan “random sample of cases en klik op “sample”. Daarna verschijnt er:



Je kan kiezen tussen een benadering (bvb 50% van alle gevallen) door onder "approximately" het percentage in te vullen. Je kan ook een exact aantal elementen selecteren. Wij gaan voor deze laatste optie en bestuderen het gedrag van steekproefgrootheden door de eerste 50 cases te nemen van het totaal. Als we dit hebben gedaan, berekenen we het rekenkundig gemiddelde en de standaardafwijking via de procedure "descriptives". Deze procedure herhalen we 50 keer. Vergeet telkens niet opnieuw de steekproefprocedure te herhalen en de filtervariabele uit te zetten of je krijgt 50 keer hetzelfde resultaat.

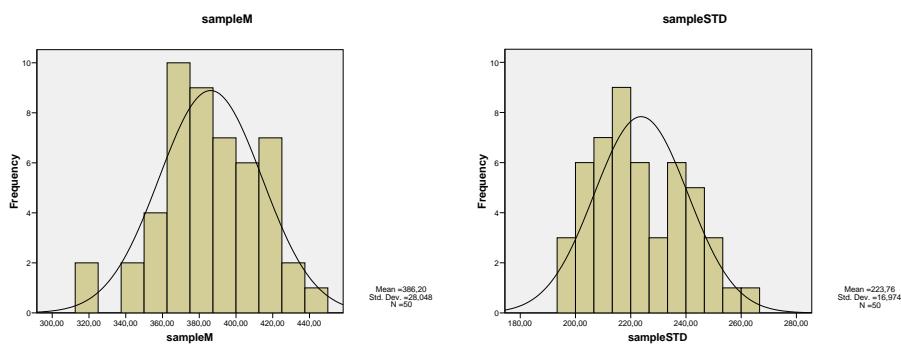
Laat ons nu een simulatie uitvoeren. We kijken naar het gedrag van steekproefgrootheden (het rekenkundig gemiddelde en de std) in 50 steekproeven van exact 50 eenheden



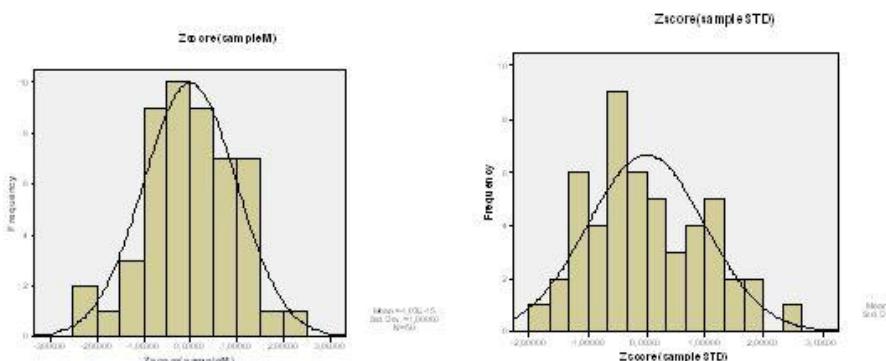
Steekproefgemiddelen en standaardafwijkingen van 50 toevalssteekproeven		Steekproefgemiddelde en standaardafwijking van de steekproefgemiddelden en standaardafwijking																																																																																																																	
<table> <tbody> <tr><td>sampleM</td><td>sampleSTD</td></tr> <tr><td>373,92</td><td>208,30</td></tr> <tr><td>410,60</td><td>217,01</td></tr> <tr><td>345,52</td><td>243,04</td></tr> <tr><td>365,50</td><td>252,23</td></tr> <tr><td>354,04</td><td>216,41</td></tr> <tr><td>393,74</td><td>207,71</td></tr> <tr><td>390,49</td><td>217,71</td></tr> <tr><td>423,57</td><td>198,77</td></tr> <tr><td>365,24</td><td>239,16</td></tr> <tr><td>404,00</td><td>211,04</td></tr> <tr><td>423,11</td><td>203,34</td></tr> <tr><td>410,26</td><td>232,19</td></tr> <tr><td>398,24</td><td>208,53</td></tr> <tr><td>318,58</td><td>205,62</td></tr> <tr><td>431,82</td><td>199,89</td></tr> <tr><td>419,68</td><td>248,75</td></tr> <tr><td>383,55</td><td>234,10</td></tr> <tr><td>383,18</td><td>245,19</td></tr> <tr><td>427,84</td><td>212,75</td></tr> <tr><td>400,24</td><td>236,39</td></tr> <tr><td>378,71</td><td>241,54</td></tr> <tr><td>358,98</td><td>218,60</td></tr> <tr><td>417,10</td><td>214,06</td></tr> <tr><td>413,08</td><td>221,41</td></tr> <tr><td>398,08</td><td>201,97</td></tr> <tr><td>447,63</td><td>204,66</td></tr> <tr><td>364,71</td><td>216,48</td></tr> <tr><td>372,06</td><td>219,46</td></tr> <tr><td>376,17</td><td>217,33</td></tr> <tr><td>397,06</td><td>224,62</td></tr> <tr><td>372,63</td><td>245,37</td></tr> <tr><td>402,35</td><td>239,67</td></tr> <tr><td>320,06</td><td>254,57</td></tr> <tr><td>374,60</td><td>250,57</td></tr> <tr><td>366,81</td><td>234,90</td></tr> <tr><td>367,48</td><td>212,63</td></tr> <tr><td>378,04</td><td>228,72</td></tr> <tr><td>386,74</td><td>221,88</td></tr> <tr><td>394,50</td><td>238,20</td></tr> <tr><td>361,38</td><td>225,25</td></tr> <tr><td>421,12</td><td>194,39</td></tr> <tr><td>390,00</td><td>223,15</td></tr> <tr><td>350,42</td><td>212,86</td></tr> <tr><td>408,91</td><td>219,38</td></tr> <tr><td>387,45</td><td>241,33</td></tr> <tr><td>376,29</td><td>203,84</td></tr> </tbody> </table>		sampleM	sampleSTD	373,92	208,30	410,60	217,01	345,52	243,04	365,50	252,23	354,04	216,41	393,74	207,71	390,49	217,71	423,57	198,77	365,24	239,16	404,00	211,04	423,11	203,34	410,26	232,19	398,24	208,53	318,58	205,62	431,82	199,89	419,68	248,75	383,55	234,10	383,18	245,19	427,84	212,75	400,24	236,39	378,71	241,54	358,98	218,60	417,10	214,06	413,08	221,41	398,08	201,97	447,63	204,66	364,71	216,48	372,06	219,46	376,17	217,33	397,06	224,62	372,63	245,37	402,35	239,67	320,06	254,57	374,60	250,57	366,81	234,90	367,48	212,63	378,04	228,72	386,74	221,88	394,50	238,20	361,38	225,25	421,12	194,39	390,00	223,15	350,42	212,86	408,91	219,38	387,45	241,33	376,29	203,84	<p style="text-align: center;">Descriptive Statistics Sample Means and STDs</p> <table border="1"> <thead> <tr> <th></th> <th>sampleM</th> <th>sampleSTD</th> </tr> </thead> <tbody> <tr> <td>N</td> <td>50</td> <td>50</td> </tr> <tr> <td>Valid</td> <td></td> <td></td> </tr> <tr> <td>Missing</td> <td>0</td> <td>0</td> </tr> <tr> <td>Mean</td> <td>386,1963</td> <td>223,7603</td> </tr> <tr> <td>Std. Deviation</td> <td>28,04839</td> <td>16,97392</td> </tr> </tbody> </table> <p style="text-align: center;">Merk op dat het gemiddelde van de steekproefgemiddelden en standaardafwijkingen een zeer goede benadering vormen van het gekende populatiegemiddelde en de standaardafwijking!</p>			sampleM	sampleSTD	N	50	50	Valid			Missing	0	0	Mean	386,1963	223,7603	Std. Deviation	28,04839	16,97392
sampleM	sampleSTD																																																																																																																		
373,92	208,30																																																																																																																		
410,60	217,01																																																																																																																		
345,52	243,04																																																																																																																		
365,50	252,23																																																																																																																		
354,04	216,41																																																																																																																		
393,74	207,71																																																																																																																		
390,49	217,71																																																																																																																		
423,57	198,77																																																																																																																		
365,24	239,16																																																																																																																		
404,00	211,04																																																																																																																		
423,11	203,34																																																																																																																		
410,26	232,19																																																																																																																		
398,24	208,53																																																																																																																		
318,58	205,62																																																																																																																		
431,82	199,89																																																																																																																		
419,68	248,75																																																																																																																		
383,55	234,10																																																																																																																		
383,18	245,19																																																																																																																		
427,84	212,75																																																																																																																		
400,24	236,39																																																																																																																		
378,71	241,54																																																																																																																		
358,98	218,60																																																																																																																		
417,10	214,06																																																																																																																		
413,08	221,41																																																																																																																		
398,08	201,97																																																																																																																		
447,63	204,66																																																																																																																		
364,71	216,48																																																																																																																		
372,06	219,46																																																																																																																		
376,17	217,33																																																																																																																		
397,06	224,62																																																																																																																		
372,63	245,37																																																																																																																		
402,35	239,67																																																																																																																		
320,06	254,57																																																																																																																		
374,60	250,57																																																																																																																		
366,81	234,90																																																																																																																		
367,48	212,63																																																																																																																		
378,04	228,72																																																																																																																		
386,74	221,88																																																																																																																		
394,50	238,20																																																																																																																		
361,38	225,25																																																																																																																		
421,12	194,39																																																																																																																		
390,00	223,15																																																																																																																		
350,42	212,86																																																																																																																		
408,91	219,38																																																																																																																		
387,45	241,33																																																																																																																		
376,29	203,84																																																																																																																		
	sampleM	sampleSTD																																																																																																																	
N	50	50																																																																																																																	
Valid																																																																																																																			
Missing	0	0																																																																																																																	
Mean	386,1963	223,7603																																																																																																																	
Std. Deviation	28,04839	16,97392																																																																																																																	

Bekijk vervolgens de histogrammen van de verdelingen en we merken op dat deze verdeling van gemiddelden en standaardafwijkingen uit de 50 steekproeven van 50 eenheden de normale verdeling begint te benaderen. Zouden we dit experiment eindeloos herhalen, dan zouden we de klokvorm nog beter benaderen. Volgens de wet van de grote getallen zouden 50 steekproeven van 100 eenheden een betere benadering vormen van de klokvorm dan 50 steekproeven van 50 eenheden.

Figuur: steekproevenverdeling van ongestandaardiseerde waarden van "sample_M" en "sample_STD"



Figuur: steekproevenverdeling van gestandaardiseerde waarden van "sample_M" en "sample_STD"



Uit deze simulatie leren we de belangrijke lessen die we nodig hebben om het principe van inferentiële statistiek te begrijpen: de kans dat we op basis van een toevalsstekproef een gemiddelde waarde uitkomen die meer dan 1.96 z-scores verwijderd ligt van het populatiegemiddelde, is uitzonderlijk klein. Dit principe zullen we toepassen bij het bestuderen van steekproefuitkomsten om hypothesen te toetsen dat gestandaardiseerde effecten significant verwijderd zijn van nul. We gebruiken de steekproefparameter en de schatting van diens standaardfout om uitspraken te doen over de populatie op basis van een steekproef.

5. TOEPASSEN INFERENTIELE STATISTIEK IN DE PRAKTIJK

In dit deel worden voorbeelden gegeven van oefeningen op de inferentiële statistiek. We werken voorbeelden uit en geven bijkomende oefeningen die tijdens de oefensessies aan bod zullen komen. We leren werken met de verdelingstafels en gaan de output van SPSS lezen om aan intervalschatting en aan hypothesesetsing te doen. We vergelijken de waarde uit de steekproef met de kritische waarde in de tabel en we leren zo de parameters van de inferentiële statistiek interpreteren.

Niet enkel univariate beschrijvende statistieken hebben een steekproevenverdeling, ook andere associatiematen zoals chi-kwadraat, regressiecoëfficiënten (richtingscoëfficiënten en determinatiecoëfficiënten) en correlatiecoëfficiënten volgen een steekproevenverdeling met een bepaald patroon. De kansen dat bepaalde waarden voorkomen is gekend vanuit de mathematische statistiek. We kunnen hierdoor gebruik maken van tabellen waarin de kansen dat een bepaalde waarde voorkomt netjes genoteerd staan. Statistische verwerkingspakketten hebben de dag van vandaag de functie van deze tabellen overgenomen.

Het gedrag van steekproefparameters ter herinnering:

Het gemiddelde, de proportie en de standaardafwijking volgen een normaalverdeling.

De richtingscoëfficiënt uit een lineaire regressieanalyse volgt een t-verdeling.

De determinatiecoëfficiënt uit een lineaire regressieanalyse volgt een F-verdeling.

De Chi-kwadraat-toets uit een contingentietabel volgt een chi-kwadraat verdeling.

Voor elke verdeling is de kans bekend dat een bepaalde waarde uitgekomen wordt.

In dit deel gaan we oefeningen maken op het schatten en toetsen. Precies omdat wiskundige statistici elk patroon in detail bestudeerd hebben, kunnen we de kansen berekenen dat we een bepaalde steekproefuitkomst vinden, onder de nulhypothese dat een verband niet bestaat in de populatie (H_0 : de richtingscoëfficiënt is 0, de determinatiecoëfficiënt is 0,...) Op basis van deze vaststellingen kunnen we nu gaan schatten en toetsen. Schatten verwijst naar het ramen van een interval van waarden waarbinnen een steekproefuitkomst naar grote waarschijnlijk ligt en toetsen betekent dat we een hypothese gaan opstellen en trachten te verwerpen. We passen de inferentiële statistiek toe om te weten of onze steekproefuitkomsten kunnen veralgemeend worden naar de populatie. Het uitgangspunt in dit hoofdstuk is het volgende: we tonen de statistische output die we eerder hebben berekend in SPSS. We hebben toen gesteld dat we ons tot de essentiële bivariate associatiematen

gingen beperken. SPSS genereert echter meer output dan nodig om een bivariate associatie naar waarde te schatten. SPSS levert “default” ook de parameters nodig voor de statistische inferentie. Deze komen hier aan bod.

De inferentie van associatiematen op nominaal niveau: jeugdbende en geslacht

Eerder vonden we dit bivariaat verband tussen het lid zijn van een jeugdbende en geslacht:

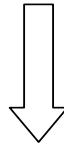
lid jeugdbende * geslacht Crosstabulation

		geslacht		Total
		meisje	jongen	
lid jeugdbende	geen lid jeugdbende	Count	786	661
		Expected Count	760,9	686,1
		% within geslacht	96,6%	90,1% 93,5%
	lid jeugdbende	Count	28	73
		Expected Count	53,1	47,9
		% within geslacht	3,4%	9,9% 6,5%
Total		Count	814	734
		Expected Count	814,0	734,0
		% within geslacht	100,0%	100,0% 100,0%

Bij afwezigheid van een statistische relatie, zijn de conditionele frequentieverdelingen identiek. Als 6.5% van de observaties lid is van een jeugdbende, dan moeten de kolompercentages voor meisjes en jongens ook allebei 6.5% bedragen. Uit de tabel blijkt echter dat dit niet zo is.

Om nu te weten of we dit verband kunnen extrapoleren naar de populatie, wordt een chi-kwadraat toets uitgevoerd. Chi-kwadraat bedraagt 26.78. Er is dus wel degelijk een verschil tussen de geobserveerde celfrequenties en de celfrequenties bij statistische onafhankelijkheid. We lezen uit de tabel af dat de p-waarde die bij de chi-kwadraat hoort, zeer klein is: de kans dat chi-kwadraat nul is in de populatie -en dus de kans dat onze resultaten op toeval berusten- is dus “verwerpelijk” klein. Bekijk een p-waarde vanuit volgende redenering: stel dat je gevraagd wordt om je hand in een zak met 95 palingen en 5 giftige slangen te steken en te zien wat er gebeurt. Je hebt 5 kansen op 100 dat het verkeerd afloopt. Die kans is niet zo groot. Sociale wetenschappers verwerpen dus een nulhypothese als ze merken dat de kans op een foutieve verwerving statistisch gezien lager ligt dan 0.05. Daarom verwerpen we in het voorbeeld de nulhypothese, omdat we weten dat de kans op een foutieve verwerving van een nulhypothese (type-1 fout) kleiner is dan 1 op 10 000.

Toetsing H₀: Chi-kwadraat = 0
in de populatie! We verwerpen
H₀ omdat p < 0.0001



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	26,785 ^b	1	,000		
Continuity Correction ^a	25,729	1	,000		
Likelihood Ratio	27,447	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	26,768	1	,000		
N of Valid Cases	1548				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 47,89.

We krijgen ook de associatiematen op nominaal niveau te zien: Phi en Cramer's V.

Symmetric Measures

	Value	Approx. Sig.
Nominal by Nominal	Phi ,132	,000
N of Valid Cases	Cramer's V ,132	,000
	1548	

- a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Hypothe-

toetsing:

H₀ = geen associatie
P < 0.0001

Is er nu een significant verband tussen geslacht en het behoren tot een jeugdbende? We lezen de informatie af uit de kruistabel en interpreteren de resultaten.

De associatiematen Cramer's V en Phi zijn hier aan elkaar gelijk omdat we een 2*2 tabel hebben. Het verband is eerder aan de zwakke kant. Cramer's V bedraagt 0.132. Chi-kwadraat bedraagt 26.78. We kunnen echter vaststellen dat we de nulhypothese (de parameter is 0 in de populatie) gerust kunnen verwerpen met een lage α -waarde ($p < 0.0001$).

De inferentie van ordinale symmetrische associatiematen

We heranalyseren de relatie tussen de uitspraken “oneerlijk zijn is ok” (antwoordcategorieën van helemaal niet akkoord tot helemaal akkoord) en “hoeveel van je vrienden hebben al iets gestolen” (geen enkele tot bijna allemaal). Deze variabelen zijn ordinaal want ze bestaan uit ordenbare antwoordcategoriën. Bijna allemaal is meer dan geen enkele, maar de afstand daartussen is niet metrisch uit te drukken. De associatie tussen ordinale kenmerken kan gebeuren aan de hand van de associatiematen Cramer's V, Gamma en de rangcorrelatiecoëfficiënt van Spearman, nl. Spearman's rho. In deze paragraaf wordt nagegaan in welke mate we de nulhypothese dat er geen associatie bestaat, kunnen verwerpen.

oneerlijk zijn is ok * hoeveel van je vrienden hebben al iets gestolen? Crosstabulation

		hoeveel van je vrienden hebben al iets gestolen?				Total	
		geen enkele	sommige	de meeste	bijna allemaal		
oneerlijk zijn is ok	helemaal oneens	Count % within hoeveel van je vrienden hebben al iets gestolen?	720 51,9%	41 28,9%	4 50,0%	0 .0%	765 49,7%
	oneens	Count % within hoeveel van je vrienden hebben al iets gestolen?	288 20,7%	34 23,9%	0 .0%	0 .0%	322 20,9%
	noch eens, noch oneens	Count % within hoeveel van je vrienden hebben al iets gestolen?	189 13,6%	24 16,9%	1 12,5%	1 50,0%	215 14,0%
	mee eens	Count % within hoeveel van je vrienden hebben al iets gestolen?	129 9,3%	24 16,9%	2 25,0%	0 .0%	155 10,1%
	helemaal mee eens	Count % within hoeveel van je vrienden hebben al iets gestolen?	62 4,5%	19 13,4%	1 12,5%	1 50,0%	83 5,4%
Total		Count % within hoeveel van je vrienden hebben al iets gestolen?	1388 100,0%	142 100,0%	8 100,0%	2 100,0%	1540 100,0%

Phi en Cramer's V zijn symmetrische maten. Dit betekent dat geen causale richting wordt verondersteld. Aangezien we te maken hebben met een r*k tabel, kijken we niet naar Phi, maar naar Cramer's V. Gamma daarentegen houdt rekening met de ordening in de data: als variabele X een hogere waarde heeft, heeft dan variabele Y ook een hogere waarde? Cramer's V bedraagt 0.111. Er is een zwakke samenhang tussen de beide variabelen. Gamma daarentegen geeft ons meer informatie dan Cramer's V. Omdat beide variabelen ordinaal zijn, nemen we best gamma. Immers, Cramer's V houdt geen rekening met de ordening in de data. Gamma zegt ons dat de associatie tussen beide kenmerken matig en positief samenhangt (Gamma= 0.375) en de significantietoets zegt ons dat we met een zeer grote zekerheid ($p < 0.0001$) de nulhypothese van afwezigheid van associatie kunnen verwerpen.

H₀ kan verworpen worden!
De associatiematen zijn statistisch significant
P< 0.0001

**Symmetric Measures**

		Value	Asymp. Std. Err. ^a	Approx. T ^b	Approx. Sig.
Nominal by Nominal	Phi	,192			,000
Nominal	Cramer's V	,111			,000
Ordinal by Ordinal	Gamma	,375	,055	5,639	,000
N of Valid Cases		1540			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Tot slot bespreken we de inferentie van de rangcorrelatiecoëfficiënt Rho. Spearman's Rho is afgeleid van de Pearson's product-moment correlatiecoëfficiënt voor interval- en ratio variabelen. De observaties worden eerst in een gewone rangorde (1^{ste}, 2^{de}, 3^{de}, ...) geplaatst op de beide variabelen. Daarna past men de formule voor de product-moment correlatiecoëfficiënt toe.

De significantietoets gebeurt standaard tweezijdig in SPSS.
We lezen een p-waarde van p < 0.0001 af en verwijderen H₀.

Correlations

			hoeveel van je vrienden hebben al iets gestolen?
		oneerlijk zijn is ok	oneerlijk zijn is ok
Spearman's rho	oneerlijk zijn is ok	Correlation Coefficient Sig. (2-tailed) N	1,000 ,158** 1548
	hoeveel van je vrienden hebben al iets gestolen?	Correlation Coefficient Sig. (2-tailed) N	,158** ,000 1540

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

De inferentie van de product-moment correlatiecoëfficiënt als symmetrische associatiemaat op metrisch niveau

Wanneer de correlatiecoëfficiënt een waarde aanneemt tussen -0.10 en +0.10 dan is dat verband verwaarloosbaar. Verbanden met een waarde tussen 0.10-0.30 (in absolute waarden) zijn zwakke tot matige bivariate verbanden, verbanden tussen 0.30-0.60 zijn matige tot sterke bivariate verbanden en verbanden die een waarde hebben die hoger is dan 0.60 zijn heel sterke bivariate verbanden. We herbekijken de samenhang tussen het aantal verschillende delicten gepleegd en het aantal verschillende delicten waarvan men slachtoffer is geworden en we zien een aanzienlijke samenhang. Kunnen we het verband extrapoleren naar de bevolking?

We verwerpen H0 omdat
de tweezijdige toets een
zeer lage p-waarde
heeft.

		aantal verschillende delicten slachtoffer	aantalverschill ende delicten gepleegd
aantal verschillende delicten slachtoffer	Pearson Correlation	1	.444**
	Sig. (2-tailed)		,000
	N	1491	1453
aantalverschillende delicten gepleegd	Pearson Correlation	,444**	1
	Sig. (2-tailed)	,000	
	N	1453	1504

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

De inferentie van regressiecoëfficiënten

Een zeer belangrijke tool in de criminologie is de lineaire regressieanalyse waar de onderzoeker geïnteresseerd is in het voorspellen van de waarden op de afhankelijke variabele op basis van een onafhankelijke variabele. We leren in deze paragraaf de parameters van regressiecoëfficiënten te interpreteren in het licht van de probleemstelling eerder behandeld tijdens de praktische oefening op de inhoudelijke interpretatie van de regressiecoëfficiënten.

We gaan er even opnieuw vanuit dat we het plegen van delicten ("kwaad doen") zien als onafhankelijke variabele. Wie kwaad doet, zou dan meer slachtoffer kunnen worden.

De inferentie van de determinatiecoëfficiënt en de correlatiecoëfficiënt

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,444 ^a	,197	,196	1,84028

- a. Predictors: (Constant), aantal verschillende delicten gepleegd
- b. Dependent Variable: aantal verschillende delicten slachtoffer

R Square is de determinatiecoëfficiënt. Het is de verklaarde variantie in Y die gebeurt op basis van X: 19.7% van de geobserveerde verschillen op de afhankelijke variabele (slachtofferschap) kan verklaard worden vanuit het plegen van delicten. Hoe weten we nu of we dit resultaat als statistisch significant kunnen beschouwen? Hiervoor moeten we de ANOVA-tabel bestuderen. De ANOVA tabel bevat de regression sum of squares, de residual sum of squares en de total sum of squares.

We herhalen even dat de regression sum of squares de som is van het kwadraat van het verschil tussen de voorspelde waarde van Y op basis van X en de gemiddelde waarde van Y.

De total sum of squares is de som van het kwadraat van het verschil tussen de geobserveerde waarde en de gemiddelde waarde.

De residual sum of squares is de som van het kwadraat van alle residuele termen. Delen we de residual sum of squares door de total sum of squares, krijgen we de aliënatiecoëfficiënt, of de niet-verklaarde variantie. I.e. als 20 procent van de variantie in Y verklaard wordt door X, dan wordt 80 procent niet verklaard door X, maar door andere onafhankelijke variabelen.

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

Regression sum of squares= teller

Total sum of squares= noemer

teller/noemer= R square

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1203,779	1	1203,779	355,451	
	Residual	4913,997	1451	3,387		
	Total	6117,776	1452			,000 ^a

a. Predictors: (Constant), aantal verschillende delicten gepleegd

b. Dependent Variable: aantal verschillende delicten slachtoffer



F-waarde: volgt F-verdeling.

P-waarde (significantieniveau)= kans om H0 foutief te verwerpen. In dit geval zijn de regression sum of squares significant verschillend van 0

In deze analyse hebben we 1452 vrijheidsgraden ($n-1$). 1 vrijheidsgraad in de teller en 1451 in de noemer. Nu begrijpen we meteen waarom een F-verdeling gebruikt wordt. De determinatiecoëfficiënt is immers een breuk (zie het theoretische gedeelte van deze syllabus). Uit de F-tabel kunnen we de kritische F-waarde opzoeken waarop we H0 verwerpen op een niveau van $\alpha = .05$. Deze kritische F-waarde blijkt 3.84 te zijn. Onze F-waarde is veel hoger. SPSS toont de F-waarde en het exacte significantieniveau. De p-waarde

is niet 0.05 maar 0.000. Hiermee is de kans op een foutief verwijderen van H₀ praktisch verwaarloosbaar.

Dus: $1203.779/6117.776 = 0.197$ of 19.7% van de waargenomen verschillen in slachtofferschap kan verklaard worden door de waargenomen verschillen in delicten plegen.

De kolom “mean square” rapporteert de “regression sum of squares” gedeeld door het aantal vrijheidsgraden. Vervolgens moeten we de regressiecoëfficiënten en de inferentiële statistieken interpreteren. We kijken naar de significantietoetsen en de betrouwbaarheidsintervallen.

Std error: standaardfout of geschatte afwijking van de populatieparameters voor B-coëfficiënten uit de lineaire vergelijking

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant) 1,449	,061	,444	23,825	,000	1,330	1,568
	aantal verschillende delicten gepleegd ,475	,025	,444	18,853	,000	,426	,525

a. Dependent Variable: aantal verschillende delicten slachtoffer

schattingen intercept en richtingscoef.
geschatte regressielijn is =
 $\hat{Y} = 1.449 + 0.475 X$

Parameters van inferentiële statistiek:
t= t-waarde bij t-verdeling regressieparameters
Sig.= α-waarde
95% Confidence Interval: voor intercept en rico

We herhalen even de betekenis van de regressieparameters. De constante (a of β_0) of het intercept bedraagt 1.449. Dit is het verwacht aantal keer dat iemand slachtoffer wordt voor iemand die geen enkele keer een crimineel feit heeft gepleegd. De ongestandaardiseerde richtingscoëfficiënt (β_1) bedraagt 0.475. SPSS geeft deze weer met de benaming B. Dit is de verwachte toename in Y als X met een eenheid stijgt, dus: $\hat{y} = \beta_0 + \beta_1 X$ Of: $\hat{y} = 1.449 + 0.475$ (aantal verschillende delicten gepleegd).

De vraag die we ons stellen is in welke mate we nu onze regressieparameters, die afkomstig zijn uit een steekproef, kunnen veralgemenen naar de populatie. Onder de conditie van een toevalssteekproef kunnen we dat, omdat we weten hoe regressieparameters zich gedragen. SPSS vergelijkt de situatie die we gevonden hebben op basis van onze steekproef met de situatie onder H_0 . H_0 wordt dan $a = 0$ (het intercept) en $b = 0$ (de richtingscoëfficient). De vraag die we ons stellen bij hypothesetoetsing is: hoe groot is de kans dat we de steekproefgrootheden intercept $a = 1.44$ en richtingscoëfficiënt $b = 0.47$ vinden onder de conditie dat beide parameters in realiteit nul zouden bedragen? Daartoe worden de standaardfouten van de regressiecoëfficiënten geschat. Deze parameters volgen een t-verdeling. De t-verdeling, zo hebben we gezien, trekt op de normale verdeling, maar heeft langere staarten. Wanneer we een regressieparameter delen door diens standaardfout, dan vinden we de t-waarde die bij de regressieparameter hoort. Uit de tabel blijkt dat de t-waarde voor het intercept (a of β_0) gelijk is aan $(1.44/0.061) = 23.82$. De t-waarde die overeenkomt met de richtingscoëfficiënt is gelijk aan $(0.47/0.025) = 18.58$. We nemen even de tabel met de t-verdeling erbij. SPSS toetst standaard tweezijdig en geeft het exacte significantieniveau. Met een aantal vrijheidsgraden van 1452 en een kritische α -waarde van 0.05 volstaat een t-waarde van 1.96 in een tweezijdige toets. Onthoud dit getal. Als je t-waarden ziet staan in een statistisch rapport die een waarde hebben groter dan 1.96, dan heb je zeker met een statistisch significant verband ($p < 0.05$ of beter) te maken. Op het examen kunnen we je een onvolledige output geven waarbij je zelf dient na te gaan of een verband significant is. Inzicht in t-waarden maakt het mogelijk zelf resultaten uit statistische studies naar waarde te schatten. In hoogstaande criminologische vaktijdschriften worden vaak enkel de parameters, de standaardfout en t-waarden weergegeven.

Tot slot willen we weten binnen welke grenswaarden onze schattingen liggen, met een bepaalde zekerheid, SPSS toetst doorgaans met een zekerheid van 95%. Daartoe kijken we naar de betrouwbaarheidsintervallen van onze regressieparameters. Een betrouwbaarheidsinterval is een schatting $+/-$ een foutenmarge. Intervalschatting komt in het voorbeeld dus neer op het schatten van een interval dat ligt rond het geschatte intercept en de richtingscoëfficiënt. We stellen het volgende vast: we weten dat we met 95% zekerheid kunnen zeggen dat het intercept tussen 1.3 als ondergrens en 1.5 als bovengrens ligt en dat de richtingscoëfficiënt tussen 0.42 als ondergrens en 0.52 als bovengrens ligt.

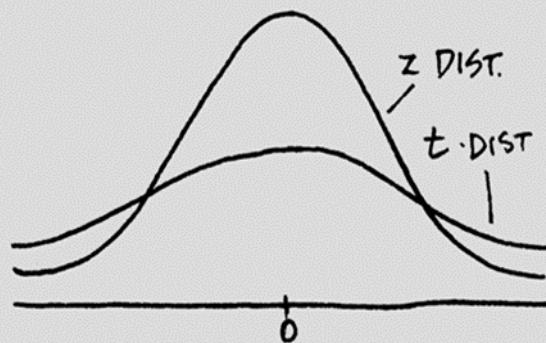
YOU CAN THINK OF THE RANDOM VARIABLE t AS THE BEST WE CAN DO UNDER THE CIRCUMSTANCES. ITS DISTRIBUTION IS CALLED STUDENT'S t , BECAUSE ITS INVENTOR, WILLIAM GOSSET, PUBLISHED UNDER THE PSEUDONYM "STUDENT."



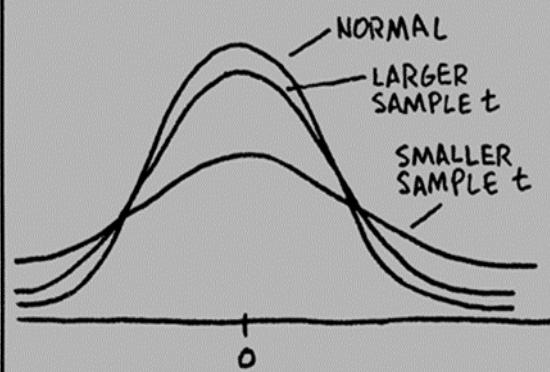
MAKING THE ASSUMPTION THAT THE ORIGINAL POPULATION DISTRIBUTION WAS NORMAL, OR NEARLY NORMAL, "STUDENT" WAS ABLE TO CONCLUDE:



t IS MORE SPREAD OUT THAN z . IT'S "FLATTER" THAN NORMAL. THIS IS BECAUSE THE USE OF s INTRODUCES MORE UNCERTAINTY, MAKING t "SLOPPIER" THAN z .



THE AMOUNT OF SPREAD DEPENDS ON THE SAMPLE SIZE. THE GREATER THE SAMPLE SIZE, THE MORE CONFIDENT WE CAN BE THAT s IS NEAR σ , AND THE CLOSER t GETS TO z , THE NORMAL.



GOSSET WAS ABLE TO COMPUTE TABLES OF t FOR VARIOUS SAMPLE SIZES

TABLE III. APPROXIMATE VALUES OF THE DISTRIBUTION OF STUDENT'S t FOR VARIOUS SAMPLE SIZES

	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100								
1	1.88	1.98	2.02	2.05	2.07	2.09	2.10	2.11	2.12	2.13	2.14	2.15	2.16	2.17	2.18	2.19	2.20	2.21	2.22	2.23	2.24	2.25	2.26	2.27	2.28	2.29	2.30	2.31	2.32	2.33	2.34	2.35	2.36	2.37	2.38	2.39	2.40	2.41	2.42	2.43	2.44	2.45	2.46	2.47	2.48	2.49	2.50	2.51	2.52	2.53	2.54	2.55	2.56	2.57	2.58	2.59	2.60	2.61	2.62	2.63	2.64	2.65	2.66	2.67	2.68	2.69	2.70	2.71	2.72	2.73	2.74	2.75	2.76	2.77	2.78	2.79	2.80	2.81	2.82	2.83	2.84	2.85	2.86	2.87	2.88	2.89	2.90	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99	3.00
2	2.22	2.26	2.28	2.30	2.31	2.32	2.33	2.34	2.35	2.36	2.37	2.38	2.39	2.40	2.41	2.42	2.43	2.44	2.45	2.46	2.47	2.48	2.49	2.50	2.51	2.52	2.53	2.54	2.55	2.56	2.57	2.58	2.59	2.60	2.61	2.62	2.63	2.64	2.65	2.66	2.67	2.68	2.69	2.70	2.71	2.72	2.73	2.74	2.75	2.76	2.77	2.78	2.79	2.80	2.81	2.82	2.83	2.84	2.85	2.86	2.87	2.88	2.89	2.90	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99	3.00																							
3	2.41	2.44	2.46	2.48	2.50	2.51	2.52	2.53	2.54	2.55	2.56	2.57	2.58	2.59	2.60	2.61	2.62	2.63	2.64	2.65	2.66	2.67	2.68	2.69	2.70	2.71	2.72	2.73	2.74	2.75	2.76	2.77	2.78	2.79	2.80	2.81	2.82	2.83	2.84	2.85	2.86	2.87	2.88	2.89	2.90	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99	3.00																																										
4	2.57	2.59	2.61	2.63	2.64	2.65	2.66	2.67	2.68	2.69	2.70	2.71	2.72	2.73	2.74	2.75	2.76	2.77	2.78	2.79	2.80	2.81	2.82	2.83	2.84	2.85	2.86	2.87	2.88	2.89	2.90	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99	3.00																																																								
5	2.77	2.78	2.79	2.80	2.81	2.82	2.83	2.84	2.85	2.86	2.87	2.88	2.89	2.90	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99	3.00																																																																									
6	2.87	2.88	2.89	2.90	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99	3.00																																																																																			
7	2.99	3.00																																																																																															
8	3.00																																																																																																
9	3.00																																																																																																
10	3.00																																																																																																
11	3.00																																																																																																
12	3.00																																																																																																
13	3.00																																																																																																
14	3.00																																																																																																
15	3.00																																																																																																
16	3.00																																																																																																
17	3.00																																																																																																
18	3.00																																																																																																
19	3.00																																																																																																
20	3.00																																																																																																
21	3.00																																																																																																
22	3.00																																																																																																
23	3.00																																																																																																
24	3.00																																																																																																
25	3.00																																																																																																
26	3.00																																																																																																
27	3.00																																																																																																
28	3.00																																																																																																
29	3.00																																																																																																
30	3.00																																																																																																
31	3.00																																																																																																
32	3.00																																																																																																
33	3.00																																																																																																
34	3.00																																																																																																
35	3.00																																																																																																
36	3.00																																																																																																
37	3.00																																																																																																
38	3.00																																																																																																
39	3.00																																																																																																
40	3.00																																																																																																
41	3.00																																																																																																
42	3.00																																																																																																
43	3.00																																																																																																
44	3.00																																																																																																
45	3.00																																																																																																
46	3.00																																																																																																
47	3.00																																																																																																
48	3.00																																																																																																
49	3.00																																																																																																
50	3.00																																																																																																
51	3.00																																																																																																
52	3.00																																																																																																
53	3.00																																																																																																
54	3.00																																																																																																
55	3.00																																																																																																
56	3.00																																																																																																
57	3.00																																																																																																
58	3.00																																																																																																
59	3.00																																																																																																
60	3.00																																																																																																
61	3.00																																																																																																
62	3.00																																																																																																
63	3.00																																																																																																
64	3.00																																																																																																
65	3.00																																																																																																
66	3.00																																																																																																
67	3.00																																																																																																
68	3.00																																																																																																
69	3.00																																																																																																
70	3.00																																																																																																
71	3.00																																																																																																
72	3.00																																																																																																
73	3.00																																																																																																
74	3.00																																																																																																
75	3.00																																																																																																
76	3.00																																																																																																
77	3.00																																																																																																
78	3.00																																																																																																
79	3.00																																																																																																
80	3.00																																																																																																
81	3.00																																																																																																
82	3.00																																																																																																
83	3.00																																																																																																
84	3.00																																																																																																
85	3.00																																																																																																
86	3.00																																																																																																
87	3.00																																																																																																
88	3.00																																																																																																
89	3.00																																																																																																
90	3.00																																																																																																
91	3.00																																																																																																
92	3.00																																																																																																
93	3.00																																																																																																
94	3.00																																																																																																
95	3.00																																																																																																
96	3.00																																																																																																
97	3.00																																																																																																
98	3.00																																																																																																
99	3.00																																																																																																
100	3.00																																																																																																



6. OEFENINGEN

- Wat is het verschil tussen een steekproefverdeling - populatieverdeling - steekproevenverdeling ?

Steekproefverdeling	
Populatieverdeling	
Steekproevenverdeling	

Wat is het verschil tussen de 'standaardafwijking' en de 'standaardfout' ?

Standaardafwijking	
Standaardfout	

Waarom is de centrale limietstelling belangrijk voor de inferentiële statistiek?

2. Zijn volgende uitspraken waar of vals?

Uitspraken	Waar	Vals
Een type-I fout betekent dat je besluit dat er een verband is, terwijl dit in realiteit niet zo is.		
Strikt genomen is het zo dat inferentiële statistiek niet dient te worden toegepast indien we populatiegegevens hebben.		
Intervalschatting toepassen is belangrijk om inzicht te krijgen in de nauwkeurigheid van een schatter		
Een p-waarde van 0.15 noemen we randsignificant		
De grootte van de steekproef is niet van invloed op de significantie van een verband.		
Bij grote steekproeven zal het steekproefgemiddelde een betere representatie zijn voor het populatiegemiddelde.		
Als een F-toets uit een bivariate regressie significant is, dan is het rekenkundig gemiddelde van Y een even goede voorspeller van Y dan de variabele X.		
Met een 95% betrouwbaarheidsinterval is het interval groter dan met een 99% betrouwbaarheidsinterval.		
Met een kleine steekproef zal het betrouwbaarheidsinterval kleiner worden.		

3. De Chi²-toets is één van de meest gebruikte manieren om relaties tussen twee of meer categorische variabelen te bestuderen. We hernemen oefening 1 van Hoofdstuk 4. Onderzoekers verzamelden gegevens over rookstatus en de diagnose longkanker bij een willekeurige steekproef van volwassenen. Elk van deze variabelen is dichotoom: een persoon rookt momenteel of niet en heeft een longkankerdiagnose of niet.

<i>Longkanker diagnose</i>			
Rookstatus	Diagnose	Geen diagnose	Totaal
Roker	60	300	
Niet-roker	10	390	
			<i>N=</i>

- Chi² = (zie oefeningen Hoofdstuk 4 – oefening 1)
- Bepaal de H₀ en H_a

H₀=

H_a =

- Om een Chi-kwadraat statistiek te interpreteren, moet je het aantal vrijheidsgraden kennen. Elke Chi-kwadraatverdeling heeft een ander aantal vrijheidsgraden en dus verschillende kritische waarden. Voor een eenvoudige Chi-kwadraat-test zijn de vrijheidsgraden $(r - 1)(k - 1)$, dat wil zeggen (het aantal rijen min 1) maal (het aantal kolommen min 1). Voor een 2×2 tabel zijn de vrijheidsgraden $(2 - 1)(2 - 1)$, ofwel 1; voor een 3×5 tabel zijn ze $(3 - 1)(5 - 1)$, of 8.
- Hoeveel bedraagt het aantal vrijheidsgraden in bovenstaande oefening?
- Hoeveel bedraagt de kritieke waarde voor $\alpha = 0.05$?
- Is er evidentie voor het verwerpen van H₀?
- Wat kunnen we besluiten over het verband tussen rookstatus en longkanker diagnose?

4. Vervolledig volgende output uit een regressieanalyse met de criminaliteitsgraad voor geweld als afhankelijke variabele en het percentage alleenstaanden als onafhankelijke variabele.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	89171,976				
	Residual	220416,710	40			
	Total	309588,686	41			

a Predictors: (Constant), % alleenstaanden

b Dependent Variable: opzet slagen 01 per km2

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	95% Confidence Interval for B		
	B	Std. Error				Lower Bound	Upper Bound	
1	(Constant)	-91,165	41,235		-2,211	,033	-174,504	-7,826
	% alleenstaanden	3,677	,914	,537			1,830	5,524

a. Dependent Variable: opzet slagen 01 per km2

- Hoe groot is het aantal vrijheidsgraden voor de regression sum of squares?
- Hoe groot is de Mean Square voor de regression, residual en total sum of squares?
- Welke F-waarde correspondeert met deze analyse en is de uitkomst statistisch significant?
- Welke t-waarde correspondeert met de richtingscoëfficiënt? En is deze significant?
Waaruit kan je dat nog opmaken?

5. Hierna volgen de scores van een leesvaardigheidstoets van een steekproef van 44 achtjarigen. Ga ervan uit dat de verdeling van de scores een normale verdeling volgt.

40 26 39 14 42 18 25 43 46 27 19
47 19 26 35 34 15 44 40 38 31 46
52 25 35 35 33 29 34 41 49 28 52
47 35 48 22 33 41 51 27 14 54 45

Veronderstel dat de standaardafwijking van de populatie leesvaardigheidsscores bekend is en gelijk is aan $\sigma = 11$. Geef een 95% betrouwbaarheidsinterval voor de verwachte score in de populatie. Rond af tot 2 cijfers na de komma.

6. Is er een statistisch significant verband tussen de beleidsverklaringen van de Premier en die van de minister van BZ?

Dave neemt een willekeurige steekproef uit de toespraken, interviews en officiële verklaringen die de Premier en de minister van Binnenlandse Zaken in de loop van een jaar hebben gegeven, en waarin wordt gerefereerd naar "gevangenisbeleid". Hij analyseert de inhoud van zijn steekproef en ontdekt vijf verschillende soorten rechtvaardigingen voor het gevangenisbeleid van de regering. Dave registreert vervolgens elke keer dat de Premier of minister van Binnenlandse Zaken naar één van de vijf rechtvaardigingstypes verwijst.

De resultaten zijn als volgt:

	Politiek mandaat		Totaal
	Premier	Minister van BZ	
Rechtvaardigingstype			
Opsluiting (beveiligen van de samenleving)	6	16	
Specifieke afschrikking	2	14	
Algemene afschrikking	4	20	
Rehabilitatie	0	15	
Gedwongen betaling	13	10	

A. Bepaal de H₀ en H_a

H ₀	
H _a	

- B. Is er een statistisch significant verband tussen de beleidsverklaringen van de Premier en die van de minister van Binnenlandse Zaken? Gebruik een significantieniveau van 5%.
- C. Zou je antwoord anders zijn als een significantieniveau van 1% werd gebruikt

7. Wietkwekers beplante 15 percelen met een nieuwe wietvariëteit. De opbrengsten van die percelen in bushels per acre zijn:

138.0 139.1 113.0 132.5 140.7 109.7 118.9 134.8

109.6 127.3 115.6 130.4 130.2 111.7 105.5

Neem aan dat $\sigma = 10$ bushels per acre.

- a. Bepaal het 90% betrouwbaarheidsinterval voor de verwachte opbrengst μ van deze wietvariëteit.
 - b. Bepaal het 95% betrouwbaarheidsinterval.
 - c. Bepaal het 99% betrouwbaarheidsinterval.
 - d. Hoe veranderen de foutmarges in (a), (b) en (c) terwijl het betrouwbaarheidsniveau toeneemt ?

8. Veronderstel dat de wietkwekers dezelfde gemiddelde x -waarde hadden gevonden in een steekproef van 60 percelen in plaats van 15 percelen.
- Bereken het 95% betrouwbaarheidsinterval voor de verwachte opbrengst μ .
 - Is de foutmarge groter of kleiner dan de foutmarge die je vond voor de steekproef van 15 percelen ? Leg in je eigen woorden uit waarom de wijziging plaats heeft.
 - Zullen de 90% en 99% intervallen voor de steekproefomvang 60 breder of smaller zijn dan die voor $n = 15$? Je hoeft die intervallen niet te berekenen maar argumenteer jouw antwoord.
 - Welke steekproefomvang is vereist om de verwachte opbrengst met 90% zekerheid te schatten tot op 4 bushels per acre ?

9. In een steekproevenverdeling van de gemiddelde leeftijd uit 1899 heeft 95% van alle steekproeven (elk bestaande uit 1000 personen) een gemiddelde leeftijd die ligt tussen 25.8 en 28.4.

Bepaal het 90% en 99% betrouwbaarheidsinterval voor μ in dezelfde steekproevenverdeling.

10. Uit de studentenbarometer, afgenoem bij 3709 studenten, blijkt dat de gemiddelde score op het einde van het 6^{de} middelbaar 72 op 100 bedraagt. De standaardafwijking in de populatie bedraagt 8.

Wat is de standaardafwijking van de steekproevenverdeling ?

- a. 0.13
- b. 8
- c. 0.0022
- d. 0.72

11. De Rector aan een Vlaamse universiteit wil laten onderzoeken wat de impact is van toelatingsexamens bij de Faculteit Recht en Criminologie. De onderzoekers vinden dat scores op deze examens grosso modo normaal verdeeld zijn volgens $N(20.8 ; 4.8)$. Neem een gerandomiseerde steekproef van 25 leerlingen die het toelatingsexamen hebben afgelegd. Wat is het gemiddelde en wat is de standaardafwijking van de gemiddelde steekproefscoore ?

- a. Het gemiddelde is 20.8 en de standaardafwijking is 0.96.
- b. Het gemiddelde is 20.8 en de standaardafwijking is 4.8.
- c. Het gemiddelde is 20.8 en de standaardafwijking is 0.19.
- d. Het gemiddelde is 25 en de standaardafwijking is 4.8.

12. De Rector aan een Vlaamse universiteit wil laten onderzoeken wat de impact is van toelatingsexamens bij de Faculteit Recht en Criminologie. De onderzoekers vinden dat scores op deze examens grosso modo normaal verdeeld zijn volgens $N(20.8 ; 4.8)$. Hoe groot is de kans dat één enkele uit de gehele examenpopulatie willekeurig gekozen leerling een score behaalt van 23 of hoger ?

- a. 32.28%
- b. 67.72%
- c. 46.00%
- d. 23.08%

13. Je bent een medisch onderzoeker die de effecten van een vegetarisch dieet op het cholesterolgehalte bestudeert. Veronderstel dat het cholesterolgehalte voor Belgische mannen tussen 20-65 jaar normaal verdeeld is, met een gemiddelde van 210 mg/dL (mg = milligram, dL = deciliter) en een standaardafwijking van 45 mg/dL. Je bestudeert een steekproef van 40 mannen uit deze leeftijdsgroep die minstens een jaar een vegetarisch dieet hebben gevolgd en je vindt dat hun gemiddelde cholesterolgehalte 190 mg/dL is.

- Bereken de z-statistiek voor de gemiddelde score van de steekproef (opgelet! Gebruik de correcte formule).
- Waar bevindt zich het gemiddelde cholesterolgehalte voor de vegetarische steekproef ten opzichte van de totale Belgische mannelijke bevolking?

14. Natuurpunt Vlaanderen doet een onderzoek. Ze willen nagaan hoeveel bomen er gemiddeld in een bos staan in Vlaanderen. Ze weten uit vorig onderzoek dat de standaardafwijking van het aantal bomen per Vlaams bos $\sigma = 2483$ bedraagt. Uit een toevalsteekproef van 83 bossen berekent men een gemiddeld aantal bomen per bos van 8596 met een standaardafwijking van 1800.

- Geef het 95% betrouwbaarheidsinterval op voor het steekproefgemiddelde.

- Geef het 80% betrouwbaarheidsinterval op voor dat steekproefgemiddelde.

- Zonder te berekenen:
 - zal het 85% betrouwbaarheidsinterval smaller of breder zijn dan het 80% betrouwbaarheidsinterval ?

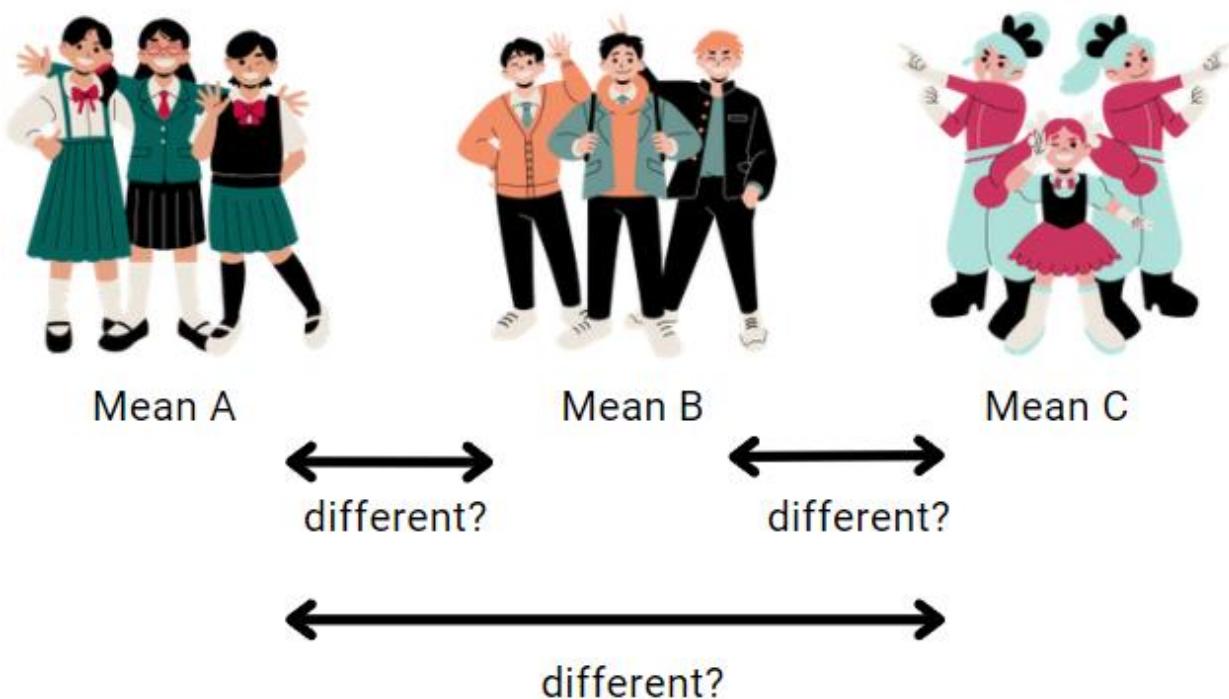
 - en smaller of breder dan het 95% betrouwbaarheidsinterval ?

- indien Natuurpunt maximum een foutmarge van 200 gewild had in hun 95% betrouwbaarheidsinterval, hoeveel bossen hadden ze dan in de steekproef moeten betrekken ?

15. De Vlaams minister-president vraagt je om een onderzoek in te stellen naar de topsalarissen bij enkele vooraanstaande banken. Uit een vooronderzoek blijkt dat de standaardafwijking € 9000 bedraagt. Op uitdrukkelijk verzoek van het ministerie mag de foutenmarge van het 95%-betrouwbaarheidsinterval rond het steekproefgemiddelde niet meer dan € 400 bedragen. Hoe groot moet je steekproef dan minstens zijn?

HOOFDSTUK 7

VARIANTIE ANALYSE



HOOFDSTUK VII

VARIANTIEANALYSE

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk zijn studenten in staat handmatig een eenvoudige variantie analyse uit te voeren. Studenten kunnen een F-toets uitvoeren en nagaan of waargenomen verschillen al dan niet significant zijn. Studenten kunnen eta-kwadraat berekenen en correct interpreteren.

2. TE ONTHOUDEN KERNBEGRIPPEN

ANOVA (<i>Analysis of variance</i>)	Een toets voor de relatie tussen een nominale en een metrische variabele. De berekeningswijze is gebaseerd op de varianties in steekproeven.
Binnengroepsvariantie (<i>within-groups</i>)	Som van de binnengroepsvarianties delen door het aantal vrijheidsgraden (= n – aantal groepen)
Eta-kwadraat	De verhouding tussen de tussengroepsvariatie en de totale variatie in Y Equivalent voor de determinatiecoëfficiënt uit een regressie analyse % van de variatie in Y dat verklaard kan worden door X
F-ratio	Tussengroepsvariantie delen door binnengroepsvariantie Hoe groter de F-waarde, hoe groter de verschillen tussen de groepen in verhouding tot de verschillen binnen de groepen
Tussengroepsvariantie (<i>between-groups</i>)	Groepsomvang * (som van de gekwadrateerde afwijkingen van de groepsgemiddelden tov het algemene gemiddelde) delen door aantal vrijheidsgraden (=aantal groepen MIN 1)

3. STATISTISCHE SYMBOLEN EN FORMULES

F-waarde = tussengroepsvariantie / binnengroepsvariantie
Verklaarde variantie Between SS / DF F = ----- = ----- Niet-verklaarde variantie Within SS / DF

4. OEFENINGEN

1. In onderstaande tabel zie je het aantal geregistreerde fietsdiefstallen in 30 Belgische gemeenten in 2008, opgesplitst naar gemeentetype.
 - Is er een verband tussen het gemeentetype en het aantal fietsdiefstallen ?
 - Hoe sterk is het verband ?
 - Is dit verband significant ?
 - Hoeveel bedraagt de goodness of fit maat ?

Grootsteden	Middelgrote steden	Rurale gemeenten
3500	1850	400
2700	1650	450
2900	1450	500
3200	1600	550
3150	1550	390
3300	1800	530
2650	1400	410
4000	1750	440
3500	1250	570
3000	1500	600

Vul de Beschrijvende statistieken en ANOVA-tabel hieronder verder aan

Werkwijze

- Bereken het gemiddelde voor elke groep.
Wat merk je reeds op ?
- Bereken de binnengroepsvariatie voor elke groep.
- Bereken de totale binnengroepsvariatie (SSwithin).
- Bepaal het aantal vrijheidsgraden. (Df_within)
- Bereken de totale binnengroepsvariantie (Mean Square within)
- Bereken de tussengroepsvariatie (SSbetween)
- Bepaal het aantal vrijheidsgraden (Df_between)
- Bereken de tussengroepsvariantie (Mean Square between)
- Bepaal de F-ratio
- Vergelijk de bekomen F-waarde met de kritieke F-waarde, gegeven het aantal vrijheidsgraden in teller en noemer en een significantieniveau $\alpha=.05$.

BESCHRIJVENDE STATISTIEKEN		
GEREGISTREERDE FIETSDIEFSTALLEN	N	Mean
1 STADSSCHOOL		
2 GEMEENTESCHOOL		
3 PLATTELANDSSCHOOL		
Total		

ANOVA					
GEREGISTREERDE FIETSDIEFSTALLEN					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

2. Controlevragen:

- Welke hypothesen kun je met een variantieanalyse testen ?
- Welke toetsingsgroothed gebruik je bij de variantieanalyse ?
- Wat gebeurt er met de F -waarde als de tussengroepsvariantie groter is dan de binnengroepsvariantie ?
- Wat gebeurt er met de p -waarde naarmate de F -waarde groter is ? En wat betekent dit voor de conclusie ?
- Wanneer gebruik je bij regressieanalyse een F -toets ?
- Wat betekent het als je $p < 0.05$ vindt bij deze F -toets ?
- Welke toets wordt nog meer bij regressieanalyse gebruikt ?

3. Twintig scholieren (tien jongens en tien meisjes) worden gemeten op een aantal variabelen zoals duur van afwezigheid gedurende afgelopen jaar en testscores bij het begin van het schooljaar. Je vindt hieronder de gegevens.

Biologisch geslacht Code 1= jongen Code 2 = meisje	Duur afwezigheid (uitgedrukt in aantal dagen)	Testscores
1	24	13
1	20	16
1	8	7
1	12	30
1	5	5
1	24	10
1	0	9
1	8	15
1	20	18
1	24	20
2	7	18
2	30	14
2	2	20
2	10	9
2	18	13
2	9	13
2	20	10
2	10	16
2	15	5
2	8	7

Voer een variantieanalyse uit tussen jongens en meisjes op 'duur afwezigheid' en op 'testscores'.

- Bepaal de F -ratio voor beide variantieanalyses
- Vergelijk de bekomen F -waarde met de kritieke F -waarde, gegeven het aantal vrijheidsgraden in teller en noemer en een significantieniveau $\alpha=.05$.
- In geval van een significant resultaat, bepaal de waarde van de goodness of fit maat.
- Interpreteer de resultaten.

BESCHRIJVENDE STATISTIEKEN		
DUUR AFWEZIGHEID		
	N	Mean
1 JONGENS		
2 MEISJES		
Total		
ANOVA		
DUUR AFWEZIGHEID		
	Sum of Squares	df
Between Groups		
Within Groups		
Total		

BESCHRIJVENDE STATISTIEKEN		
TESTSCORES		
	N	Mean
1 JONGENS		
2 MEISJES		
Total		
ANOVA		
TESTSCORES		
	Sum of Squares	df
Between Groups		
Within Groups		
Total		

4. Zesendertig personen namen deel aan een experiment om de effecten van alcohol op het rijvermogen te ontdekken. Ze werden willekeurig toegewezen aan drie verschillende condities: placebo (geen alcohol), lage alcohol en hoge alcohol. Het niet-alcoholische drankje zag er precies hetzelfde uit en smaakte hetzelfde als de andere drankjes(!). Deelnemers werden gewogen en kregen de passende hoeveelheid drank. Na een half uur drinken reden de deelnemers tien minuten in een simulator, en het aantal gemaakte fouten werd automatisch geregistreerd door de computer. De gegevens staan vermeld in onderstaande tabel.

H_0 = Er is geen verband tussen alcohol en rijvermogen.

H_a = Er is een verband tussen alcohol en rijvermogen.

Observaties	Placebo	Lage alcohol	Hoge alcohol
1	5	5	8
2	10	7	10
3	7	9	8
4	3	8	9
5	5	2	11
6	7	5	15
7	11	6	7
8	2	6	11
9	3	4	8
10	5	4	8
11	6	8	17
12	6	10	11

- Voer een variantie-analyse uit.
- Kunnen we de nulhypothese verwerpen dat er geen verband is tussen leeftijd en delict-type?

ANOVA

RIJVERMOGEN

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

- Hoeveel bedraagt de goodness of fit maat? Hoe sterk is het verband?

5. In onderstaande tabel vind je de leeftijd van 30 veroordeelde daders van witteboordencriminaliteit. De veroordeelden werden ingedeeld volgens delict-type: fraude – omkoping – witwassen. We willen nagaan of de drie groepen statistisch significant verschillen naar gemiddelde leeftijd. Met andere woorden: is er een verband tussen leeftijd en delict-type?

H_0 = Veroordeelde daders van fraude, omkoping en witwassen verschillen niet statistisch significant naar gemiddelde leeftijd (er is geen verband tussen leeftijd en delict-type)

H_a = Veroordeelde daders van fraude, omkoping en witwassen verschillen statistisch significant naar gemiddelde leeftijd (er is een verband tussen leeftijd en delicttype)

OBSERVATIES	FRAUDE	OMKOPING	WITWASSEN
1	19	28	35
2	21	29	46
3	23	32	48
4	25	40	53
5	29	42	58
6	30	48	61
7	31	58	62
8	35	58	62
9	42	64	62
10	49	68	75
N	10	10	10

- Voer een variantie-analyse uit.
- Kunnen we de nulhypothese verwijzen dat er geen verband is tussen leeftijd en delict-type?

ANOVA

LEEFTIJD

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

- Hoeveel bedraagt de goodness of fit maat? Hoe sterk is het verband?

6. Veronderstel volgende fictieve scenario:

Een groep laatstejaarsstudenten besluit om hallucinogene drugs te nemen tijdens de colleges. Aan het eind van het semester is er een examen.

De studenten die drugs gebruiken tijdens de colleges behaalden volgende cijfers (%):

23 89 62 11 76 28 45 52 71 28

De studenten die GEEN drugs gebruiken tijdens de colleges behaalden volgende cijfers (%):

45 52 68 74 55 62 58 49 42 57

- Wat is de onafhankelijke variabele in dit fictieve scenario?
- Wat is de afhankelijke variabele in dit fictieve scenario?
- Welke groep heeft de hoogste gemiddelde score?
- In welke groep is de variatie het grootst?
- Is er een statistisch significant verschil tussen beide groepen voor wat de gemiddelde score betreft?
- Als er een statistisch significant verschil is, hoe groot is het effect?

7. Heeft blootstelling aan anti-pest interventiecampagnes een impact op de kennis van jongeren over pesten op school?

In een experimenteel onderzoek wordt nagegaan of blootstelling aan verschillende anti-pest interventiecampagnes een impact heeft op kennis van jongeren over pesten op school.

De experimentele setting omvat drie groepen respondenten:

Groep 1 wordt blootgesteld aan anti-pest campagne nummer 1.

Groep 2 wordt blootgesteld aan anti-pest campagne nummer 2.

Groep 3 wordt niet blootgesteld aan een anti-pest campagne en is de controlegroep.

Elke groep telt 9 onderzoekseenheden. In de tabel hieronder vind je de score van elke respondent in elke groep op een anti-pest kennistest afgenoem na blootstelling aan de anti-pest campagne.

ONDERZOEKSVRAAG

Verschillen groepen die verschillend zijn *blootgesteld aan anti-pest-campagnes* significant van elkaar op gemiddelde *scores op een anti-pest-kennistest!*

	GROEP 1 Campagne 1	GROEP 2 Campagne 2	GROEP 3 controlegroep
1	13	12	8
2	4	14	5
3	7	14	12
4	14	15	6
5	16	19	11
6	12	23	10
7	16	14	17
8	13	21	7
9	13	16	11

OPDRACHT

Voer een variantie analyse uit en beantwoord onderstaande vragen.

ANOVA

Scores op anti-pest-kennistest

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups					
Within Groups					
Total					

- Zijn de verschillen tussen de groepen groter dan de verschillen binnen de groepen?
- Hoeveel bedraagt de *F*-statistiek?
- Hoeveel bedraagt de kritieke *F*-waarde, gegeven aantal *DF* in teller en noemer?
- Kan de nulhypothese verworpen worden dat er *geen* verschillen zijn in gemiddelde scores op een anti-pest-kennistest tussen de groepen?
- Hoeveel bedraagt Eta-squared?
- Wat betekent dit concreet? Rapporteer in eigen woorden.
- Hoe sterk is het verband tussen blootstelling aan anti-pest interventiecampagnes en scores op een anti-pest-kennistest?
 - a. Welke statistiek bereken je?
 - b. Interpreteer.

8. WAAR OF VALS

UITSPRAAK	WAAR	VALS
Associatiematen hebben in geval van nominale variabelen geen richting.		
In de samenhang tussen religie en aantal jaren opleiding is de richting van de samenhang niet van toepassing.		
Een toename in opleiding hangt samen met een toename in salaris. Dit is een voorbeeld van een negatieve samenhang.		
Phi kan gebruikt worden om de samenhang te bestuderen tussen etniciteit en politieke voorkeur.		
Eta ² wordt gebruikt wanneer de onafhankelijke variabele categorisch is en de afhankelijke ook.		
ANOVA kan enkel toegepast worden ingeval twee groepsgemiddeldes worden vergeleken.		
De binnengroepsvariatie is de totale hoeveelheid variatie van alle eenheden in een steekproef in verhouding tot de subgroepgemiddelden.		
De binnengroepsvariatie is de hoeveelheid variatie van alle eenheden in een steekproef in verhouding tot het algemene gemiddelde.		
Eta ² wordt berekend wanneer de onafhankelijke variabele van het metrische meetniveau is en de afhankelijke categorisch.		

HOOFDSTUK 8

PARTIËLE CORRELATIE

I DID A STATISTICAL ANALYSIS AND FOUND NO CORRELATION BETWEEN MY EFFORTS AND MY REWARDS.



HOOFDSTUK VIII

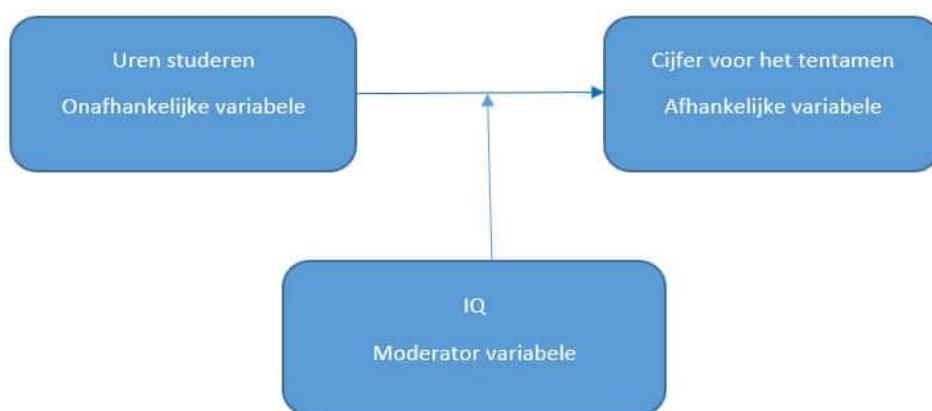
PARTIELE CORRELATIE

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk hebben studenten inzicht in de beperkingen van de bivariate statistiek en is het principe van de statistische controle goed begrepen. Inzicht in de partiële correlatie is nodig aangezien een verband tussen twee variabelen kan verzwakken, wegvalLEN of versterken na controle voor een derde variabele. Studenten zijn zelf in staat om een partiële correlatie tussen twee variabelen onder statistische controle van een derde variabele te berekenen.

2. TE ONTHOUDEN KERNBEGRIPPEN

Interactie	Er is sprake van interactie als het effect van een onafhankelijke variabele X op een afhankelijke variabele Y afhangt van de waarde van een andere onafhankelijke variabele Z
------------	---



Opgelet! Niet verwarring met 'controleren voor'
In dat geval wordt er vanuit gegaan dat het effect van x op y min of meer hetzelfde blijft als de controlerende variabele (z) andere waarden aanneemt

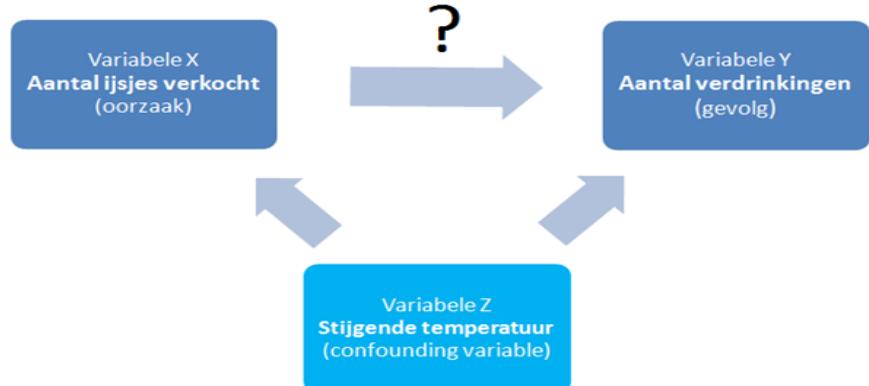
Meervoudige regressie	Statistische techniek waarbij een lineair verband wordt berekend tussen 1 afhankelijke variabele en minimum 2 onafhankelijke variabelen
-----------------------	---

Multivariate analyse	In dergelijke analyses wordt nagegaan of bivariate relaties ook blijven bestaan als rekening wordt gehouden met andere variabelen. Men spreekt in dat verband van 'controleren voor'
----------------------	--

Partiële correlatie	De correlatie tussen twee variabelen ontdaan van de invloed van extra variabelen. Een soort netto correlatie
---------------------	--

Schijnverband	Er kan sprake zijn van een schijnverband als de oorspronkelijke samenhang tussen een x-variabele en een y-variabele verdwijnt of van teken verandert na rekening te houden met een andere variabele (z)
---------------	---

Spurieuze relatie Schijnrelatie
Een bivariaat verband is het gevolg van een gemeenschappelijke oorzaak



- Statistische controle** Elke variabele die gecorreleerd is met de onafhankelijke variabele en mede bepalend kan zijn voor de score op de afhankelijke variabele is een storende variabele. Controleren voor een storende variabele betekent dat deze constant wordt gehouden en enkel wordt gekeken naar de variatie in de afhankelijke variabele en in de onafhankelijke variabele die samenhangt met eenzelfde niveau in de controlevariabele
- Suppressie-effect** Als de samenhang tussen x-variabele en y-variabele sterker wordt na controle voor één of meer andere variabele, dan is er sprake van suppressie of onderdrukking

3. STATISTISCHE SYMBOLEN EN FORMULES

partiële correlatie: verband tussen twee variabelen X en Y gecontroleerd voor een derde variabele Z

$$r_{XY} - (r_{XZ})(r_{YZ})$$

$$r_{XY \cdot Z} = \frac{r_{XY} - (r_{XZ})(r_{YZ})}{\sqrt{1 - r_{XZ}^2} \times \sqrt{1 - r_{YZ}^2}}$$

$$\sqrt{1 - r_{XZ}^2} \times \sqrt{1 - r_{YZ}^2}$$

OEFENINGEN

1. Hieronder vind je de SPSS output van de bivariate lineaire regressie waarin het aantal arrestaties na gevangenisstraf gedefinieerd wordt in functie van het aantal jaren gevangenisstraf in een steekproef van 20 ex-gedetineerden. (hypothetische data).

Descriptive Statistics

	Mean	Std. Deviation	N
arrestatiesNA	3,10	2,360	20
gevangenisstraf	3,15	,988	20

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,716 ^a	,512	,485	1,694

a. Predictors: (Constant), gevangenisstraf

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-2,283	1,295	-1,763	,095
	gevangenisstraf	1,709	,393		

a. Dependent Variable: arrestatiesNA

- Hoeveel procent van de variatie in het *aantal arrestaties na gevangenisstraf* kan verklaard worden op basis van *aantal jaren gevangenisstraf*?
- Wat is de verwachte toename in *aantal arrestaties na gevangenisstraf* met elk bijkomend jaar *gevangenisstraf*?

- Een bivariaat model voor arrestaties na gevangenisstraf dat stelt dat aantal jaren gevangenisstraf de enige impactfactor is, is echter een twijfelachtig model. Gezond verstand zegt ons immers dat dit bivariate model niet correct is gedefinieerd, er zijn immers nog andere factoren die het aantal arrestaties kunnen beïnvloeden. Als dit waar is - dat relevante factoren die verband houden met het aantal jaren gevangenisstraf uit het model zijn weggelaten, dan kan de regressiecoëfficiënt een zeer misleidende schatting geven van het effect van gevangenisstraf op arrestaties. We voegen daarom een derde variabele toe aan ons model: *aantal arrestaties voor gevangenisstraf*.

	Aantal arrestaties na gevangenisstraf	Aantal jaren gevangenisstraf	Aantal arrestaties voor gevangenisstraf
Aantal arrestaties na gevangenisstraf	1		
Aantal jaren gevangenisstraf	.72	1	
Aantal arrestaties voor gevangenisstraf	.76	.63	1

- Hoe interpreteer je de relatie tussen *aantal arrestaties voor gevangenisstraf* en *aantal arrestaties na gevangenisstraf*?
- Is *aantal arrestaties voor gevangenisstraf* derhalve een relevante factor om mee te nemen in het regressiemodel ter verklaring van *aantal arrestaties na gevangenisstraf*?

- c) Hoe weten we of de relatie tussen ‘*aantal jaren gevengenisstraf*’ en ‘*aantal arrestaties na gevengenisstraf*’ niet gewoon het gevolg is van het feit dat zij die een langere gevengenisstraf uitzitten over het algemeen al een zwaarder verleden hebben (vertaald als een *hoger aantal arrestaties voor gevengenisstraf*)?
- Wat gebeurt er met de samenhang tussen ‘*aantal jaren gevengenisstraf*’ en ‘*aantal arrestaties na gevengenisstraf*’ als we controleren voor ‘aantal arrestaties voor gevengenisstraf’?
 - Vul onderstaande tabel aan met de partiële correlatiecoëfficiënt.
 - Interpreteer de bevinding.

CORRELATIES

Controlevariabele	Aantal jaren gevangenisstraf	Aantal arrestaties NA gevengenis
Aantal arrestaties VOOR gevangenisstraf	Aantal jaren gevangenisstraf	1
	Aantal arrestaties NA gevangenisstraf	??

2. Partiële correlatie is een statistische manier om het verband tussen twee variabelen te berekenen als je de invloed van één (of meer) variabele(n) wilt verwijderen. Dit wordt het *corrigeren voor een derde variabele* genoemd. In essentie komt het erop neer dat de waarde van de correlatiecoëfficiënt wordt gecorrigeerd door rekening te houden met de invloed van een derde variabele (of meerdere variabelen).

Je kan de partiële correlatie toepassen als de zero-order correlaties beschikbaar zijn. Dit zijn de bivariate correlaties tussen de hoofdvariabelen.

Hieronder vind je een correlatiematrix met de correlaties tussen drie variabelen.

	NUMERIEKE SCORE = X	VERBALE SCORE = Y	LEEFTIJD = Z
NUMERIEKE SCORE (= X)	1.00		
VERBALE SCORE (= Y)	.97	1.00	
LEEFTIJD (= Z)	.80	.85	1.00

- Bereken XY.Z en vul onderstaande presentatie van de resultaten verder aan.

Presentatie van de resultaten

"Omdat leeftijd zowel met verbale als met numerieke vaardigheid correleert, besloten we om te onderzoeken wat het effect van een correctie voor leeftijd is op de correlatie. Na het uitvoeren van een partiële correlatie neemt de correlatiecoëfficiënt (toe/af), van (= zero-order correlatie tussen numerieke score en verbale score) naar (=waarde van de partiële correlatie). Dit is een (grote/kleine) verandering. Leeftijd heeft (weinig/veel) impact op de correlatie tussen verbale en numerieke scores."

3. Een onderzoeker vindt voor een grote stedelijke politiezone een correlatie van -.40 tussen prestaties op een fysieke bekwaamheidstest (X) en salaris (Y) in een steekproef van 50 politieagenten. Op het eerste zicht zou men hieruit kunnen besluiten dat de politiezone een lager salaris uitbetaalt aan agenten die in fysieke topconditie zijn. Het houdt echter meer steek dat het aantal dienstjaren (Z) een invloed heeft op zowel fysieke bekwaamheid als op salaris.

	FYSIEKE BEKWAAMHEID = X	SALARIS = Y	DIENSTJAREN = Z
FYSIEKE BEKWAAMHEID	1.00		
SALARIS	-.44	1.00	
DIENSTJAREN	-.68	.82	1.00

- Bereken, aan de hand van de mathematische formule, de partiële correlatiecoëfficiënt (notatie= $R_{xy.z}$ = de partiële correlatie van X en Y onder controle van Z).
- Wat is jouw besluit?

4. EXTRA OEFENING

Een onderzoeker is geïnteresseerd in de samenhang tussen werkloosheidsduur en sollicitatie-activiteit onder ex-gedetineerden. Hij interviewt 6 werkloze ex-gedetineerden en bevraagt onder meer :

- Aantal weken dat zij werkloos zijn (X)
- Het aantal sollicitaties dat zij kunnen voorleggen (Y)
- Hun leeftijd (Z).

Je vindt hieronder de resultaten.

AANTAL WEKEN WERKLOOS =X	AANTAL SOLLICITATIES =Y	LEEFTIJD =Z
2	8	30
7	3	42
5	4	36
12	2	47
1	5	29
10	2	56

- Bereken, op basis van een serie regressievergelijkingen, de partiële correlatiecoëfficiënt voor de samenhang tussen 'aantal weken werkloos' en 'het aantal sollicitaties' onder controle van 'leeftijd'.
- Wat is jouw besluit ?

HOOFDSTUK 9

REGRESSIE ANALYSE MET TWEE ONAFHANKELIJKE VARIABELEN

Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT.
 x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM

$$y = a + bx$$



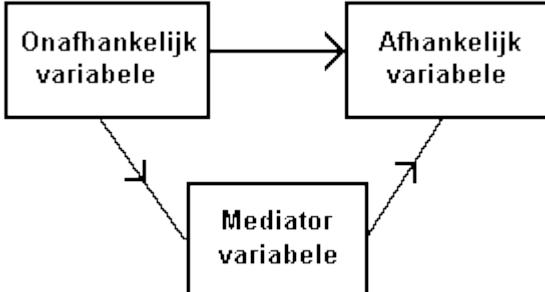
HOOFDSTUK IX

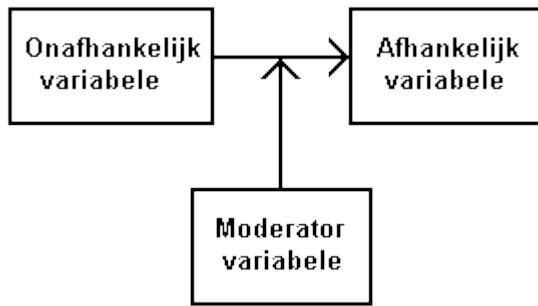
MULTIPLE REGRESSIE ANALYSE

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk begrijpen studenten de meervoudige regressieanalyse en kunnen zij deze zelf toepassen (zelf uitrekenen) in de situatie met twee onafhankelijke variabelen. Studenten begrijpen waarom niet zomaar twee afzonderlijke bivariate analyses bij elkaar kunnen opgeteld worden. Studenten begrijpen dat de lineaire OLS-regressieanalyse niet steeds bruikbaar is en dat bij schendingen van assumpties voor andere methoden dient gekozen te worden. Studenten kunnen de regressiecoëfficiënten inhoudelijk interpreteren. Tot slot kunnen studenten statistische interactie uitleggen aan de hand van een voorbeeld.

2. TE ONTHOUDEN KERNBEGRIPPEN

Confounder	Controlevariabele
Interactie-effect	De sterkte van een relatie tussen twee kenmerken wordt beïnvloed door een derde kenmerk (een moderatorvariabele)
Mediatorvariabele	Statistische variabele die de relatie tussen twee andere variabelen X en Y verklaart
	 <pre> graph LR A[Onafhankelijk variabele] --> B[Afhankelijk variabele] A -.-> C[Mediator variabele] B -.-> C </pre>
Meervoudige of multiple regressieanalyse	Statistische techniek waarbij een lineair verband wordt berekend tussen een afhankelijke variabele van het metrische niveau op basis van meerdere onafhankelijke variabelen eveneens van het metrische niveau
Moderatorvariabele	Verandert het effect dat een variabele X heeft op een variabele Y afhankelijk van de waarde van de moderatorvariabele Z



Multicollineariteit

De mate van overlap/correlatie die bestaat tussen de onafhankelijke variabelen van een regressieanalyse. De samenhang tussen de onafhankelijke variabelen moet zo klein mogelijk zijn.

3. STATISTISCHE SYMBOLEN EN FORMULES

Determinatiecoëfficiënt	$R^2 = \beta_1 r_{y1} + \beta_2 r_{y2}$
Gestandaardiseerde regressiegewichten	$\beta_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$ $\beta_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}$
Intercept	$a = \text{gemiddelde } Y - b_1 * \text{gemX1} - b_2 * \text{gemX2}$
Ongestandaardiseerde regressiegewichten	$b_1 = \left(\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - (r_{x1,x2})^2} \right) \left(\frac{SD_y}{SD_{x1}} \right)$ $b_2 = \left(\frac{r_{y,x2} - r_{y,x1} r_{x1,x2}}{1 - (r_{x1,x2})^2} \right) \left(\frac{SD_y}{SD_{x2}} \right)$
Regressievergelijking met 2 onafhankelijke variabelen	$Y = a + b_1 X_1 + b_2 X_2 + e$

4. OEFENINGEN

1. *Helpt het als studenten een privébegeleider inschakelen om een hogere eindscore op hun statistiekexamen te behalen ?*

Op basis van gegevens over 3492 studenten (van wie er 573 een privébegeleider hadden) werd het model $E(y) = a + b_1*x_1 + b_2*x_2$ geschat, waarin y = de eindscore op het examen statistiek, x_1 = score op een tussentest statistiek en x_2 = de mate waarin de student een privébegeleider had.

- Het geschatte model heeft een R^2 -waarde van 0.76. Interpreteer dit resultaat.
- De schatting voor b_2 is 19 met een standaardfout van 3. Gebruik deze informatie om een 95% betrouwbaarheidsinterval voor b_2 te construeren. Geef een interpretatie van het gevonden interval.
- Wat kun je, uitgaande van voorgaande, zeggen over het effect van een privébegeleider op de eindscore van het statistiekexamen ?

2. Een onderzoeker wil de impact van sociale bindingen (SOCBIND) en extraverte persoonlijkheid (EXTRAVERT) op tevredenheid op het werk (TEVREDEN) analyseren. In de tabel hieronder vind je de gegevens op deze drie variabelen voor 10 respondenten. De drie variabelen zijn van het metrische meetniveau.

Observaties	SOCBIND	EXTRAVERT	TEVREDEN
1	20	15	20
2	10	30	15
3	4	5	5
4	17	16	20
5	10	14	15
6	11	8	10
7	7	7	8
8	4	4	5
9	15	10	17
10	17	5	17

- a. Wat is de afhankelijke variabele? Wat zijn de onafhankelijke variabelen?
- b. Bereken met de hand de regressie-coëfficiënten (intercept, ongestandaardiseerde en gestandaardiseerde regressieparameters)
- c. Interpreteer elke regressie-coëfficiënt.
- d. Hoeveel bedraagt de totale verklaarde variantie in de afhankelijke variabele ?
- e. Welke onafhankelijke variabele heeft het sterkste relatieve netto-effect?

3. Onderstaande uitspraken over de meervoudige lineaire regressie zijn FOUT. Leg uit wat er fout is en waarom.

FOUTE UITSPRAKEN	WAT IS FOUT EN WAAROM?
<i>De meervoudige correlatie geeft de verhouding weer van de variatie in de te verklaren variabele dat is gegeven door de verklarende variabelen.</i>	
<i>Eén van de assumpties in meervoudige lineaire regressie is dat de verdeling van elke verklarende variabele normaal moet zijn.</i>	
<i>Een hoge R-kwadraat waarde betekent een causaal verband tussen de onafhankelijke en afhankelijke variabelen.</i>	
<i>Alle onafhankelijke variabelen moeten sterk gecorreleerd zijn met de afhankelijke variabele om een goed regressiemodel te hebben.</i>	

4. Veronderstel onderstaande hypothetische gegevens van 20 ex-gedetineerden.

Respondenten	Aantal arrestaties na gevangenisstraf	Aantal jaren gevangenisstraf	Aantal arrestaties voor gevangenisstraf
1	0	2	4
2	0	3	2
3	1	1	2
4	1	2	3
5	1	3	3
6	1	3	2
7	2	4	3
8	2	2	3
9	2	2	1
10	3	3	2
11	3	3	3
12	3	3	3
13	4	3	4
14	4	4	3
15	4	4	4
16	4	4	5
17	5	4	4
18	6	4	5
19	7	5	5
20	9	4	6

- Definieer een meervoudige lineaire regressie waarin *het aantal arrestaties na gevangenisstraf* wordt gedefinieerd in functie van *het aantal jaren gevangenisstraf* en *het aantal arrestaties voor gevangenisstraf*.
- Bereken de regressiecoëfficiënten.
- Interpreteer elke regressiecoëfficiënt.
- Welke onafhankelijke variabele heeft het sterkste netto-effect?
- Hoeveel bedraagt de verklaringskracht van het model op basis van de twee onafhankelijke variabelen.
- Wat is het verwachte *aantal arrestaties na gevangenisstraf* bij een gevangenisstraf van 4 jaar en 3 arrestaties voor gevangenisstraf?

5. In onderstaande tabel worden de ruwe data weergegeven van 10 atleten. Op basis van deze scores willen we predicties maken van prestaties op basis van uren training en scores op een motivatietest.

Observaties	Prestatie-scores	Uren training	Motivatie-scores
1	67	6	36
2	87	8	32
3	87	8	43
4	56	5	26
5	72	7	31
6	57	5	38
7	60	6	42
8	92	9	48
9	56	5	33
10	67	6	30

- a) Wat is de afhankelijke variabele? Wat zijn de onafhankelijke variabelen?
- b) Bereken met de hand de regressie-coëfficiënten (intercept, ongestandaardiseerde en gestandaardiseerde regressieparameters)
- c) Hoeveel bedraagt de totale verklaarde variantie in de afhankelijke variabele?
- d) Wat is de verwachte prestatie bij een training van 7 en een motivatie van 48?
- e) Wat is de training bij een verwachte prestatie van 95 en een motivatie van 45?

6. De European Social Survey (ESS) of het Europees Sociaal Onderzoek, is een cross-nationaal onderzoek dat in verschillende Europese landen beoogt attitudes, meningen en gedragspatronen van de bevolking in kaart te brengen. Sinds 2002 wordt de enquête iedere twee jaar georganiseerd. Ook België neemt hieraan deel. Aan 1704 respondenten werd gevraagd '*Hoe tevreden bent u met de democratie in uw land?*'(Y).

We voerden een meervoudige lineaire regressieanalyse uit met als onafhankelijke variabelen:

In blok 1: gender, leeftijd, opleidingsniveau

In blok 2: attitude_immigratie, vertrouwen_politieke_instellingen.

Attitude_immigratie : peilt naar de attitude van de respondent ten aanzien van immigratie. Hoge waarden verwijzen naar een positieve attitude ten aanzien van immigratie

Vertrouwen_politieke_instellingen : peilt naar de mate van vertrouwen in politieke instellingen in België bij de respondent. Hoge waarden verwijzen naar veel vertrouwen in politieke instellingen in België.

Hieronder vind je de SPSS-output. Beantwoord de onderstaande vragen.

Descriptive Statistics

	Mean	Std. Deviation	N
hoe tevreden bent u met de werking van de democratie in uw land	5,20	2,215	1.640
gender	,49	,500	1.640
leeftijd	54,0628	18,67611	1.640
Opleidingsniveau	8,13	4,359	1.640
attitude ten aanzien van immigratie	14,6884	5,33924	1.640
Vertrouwen_politieke_instellingen	12,1994	5,89235	1.640

Model Summary

Mode I	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Sig. F Change
				R Square Change	F Change	df1	df2		
1	,175 ^a	,031	,029	2,183	,031	17,206	3	1.636	,000
2	,549 ^b	,302	,300	1,854	,271	317,343	2	1.634	,000

a. Predictors: (Constant), Opleidingsniveau, gender, leeftijd

b. Predictors: (Constant), Opleidingsniveau, gender, leeftijd, vertrouwen_politieke_instellingen, attitude_immigratie

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	245,960	3	81,987	17,206	,000
	Residual	7.795,424	1.636	4,765		
	Total	8.041,385	1.639			
2	Regression	2.426,805	5	485,361	141,2535	
	Residual	5.614,579	1634	3.4361		
	Total	8.041,385	1.639			

a. Dependent Variable: hoe tevreden bent u met de werking van de democratie in uw land

b. Predictors: (Constant), Opleidingsniveau, gender, leeftijd

c. Predictors: (Constant), Opleidingsniveau, gender, leeftijd, vertrouwen_politieke_instellingen, attitude_immigratie

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficient	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	5,206	,215	24,234	,000	4,785	5,627
	gender	,301	,108	,068	2,794	,005	,090
	leeftijd	-,012	,003	-,098	-3,995	,000	-,017
	Opleidingsniveau	,058	,013	,115	4,663	,000	,034
2	(Constant)	2,335	,224	10,435	,000		
	gender	,167	,092	,038	1,819	,069	
	leeftijd	-,006	,002	-,052	-2,483	,013	
	Opleidingsniveau	-,003	,011	-,005	-,242	,809	
	attitude_immigratie	,076	,010	,183	7,864	,000	
	vertrouwen_politiek_e_instellingen	,166	,008	,441	19,637	,000	

a. Dependent Variable: hoe tevreden bent u met de werking van de democratie in uw land

- Op basis van hoeveel respondenten is de analyse uitgevoerd ?
- Hoeveel bedraagt de multiple determinatiecoëfficiënt en hoe wordt deze maat geïnterpreteerd ?
- Hoeveel bedraagt de *F*-waarde in blok 2 in de ANOVA-tabel ?

- Is de *F*-waarde statistisch significant op het niveau $\alpha = .001$? (zoek in de *F*-tabel de kritische *F*-waarde op gegeven het aantal df in teller en noemer). Interpreteer.

- Welke variabele in blok heeft het sterkste relatieve effect? Is er sprake van een statistisch significant effect? Beargumenteer.

- Hoe interpreteer je dit relatieve effect?

- Welke variabelen hebben in blok 2 geen statistisch significant effect?

- Geef voor de variabele ‘vertrouwen_politieke_instellingen’ in blok 2 het 95% betrouwbaarheidsinterval. Interpreteer.

7. In onderstaande tabellen vindt u de output van een multipele regressieanalyse met als afhankelijke variabele '*dagsalaris van een topmodel*' en als onafhankelijke variabelen '*leeftijd*', '*jaren ervaring als model*' en '*gepercipieerde schoonheid*' (mate waarin een model door een panel van experten als 'aantrekkelijk' wordt gepercipieerd).

Interpreteer de output.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics		
						F Change	df1	df2
1	,429 ^a	,184	,173	14,57213	,184	17,066	3	227

a. Predictors: (Constant), gepercipieerde schoonheid, Jaren ervaring, leeftijd

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.871,964	3	3.623,988	17,066	,000 ^b
	Residual	48.202,790		212,347		
	Total	59.074,754		230		

a. Dependent Variable: dag-salaris

b. Predictors: (Constant), gepercipieerde schoonheid, Jaren ervaring, leeftijd

Coefficients^a

Model		B	Std. Error	Standardized Coefficients		95,0% Confidence Interval for B		
				Unstandardized Coefficients				
				Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-60,890	16,497		-3,691	,000	-93,396	-28,384
	Leeftijd	6,234	1,411	,942	4,418	,000	3,454	9,015
	Jaren ervaring	-5,561	2,122	-,548	-2,621	,009	-9,743	-1,380
	Gepercipieerde schoonheid	-,196	,152	-,083	-1,289	,199	-,497	,104

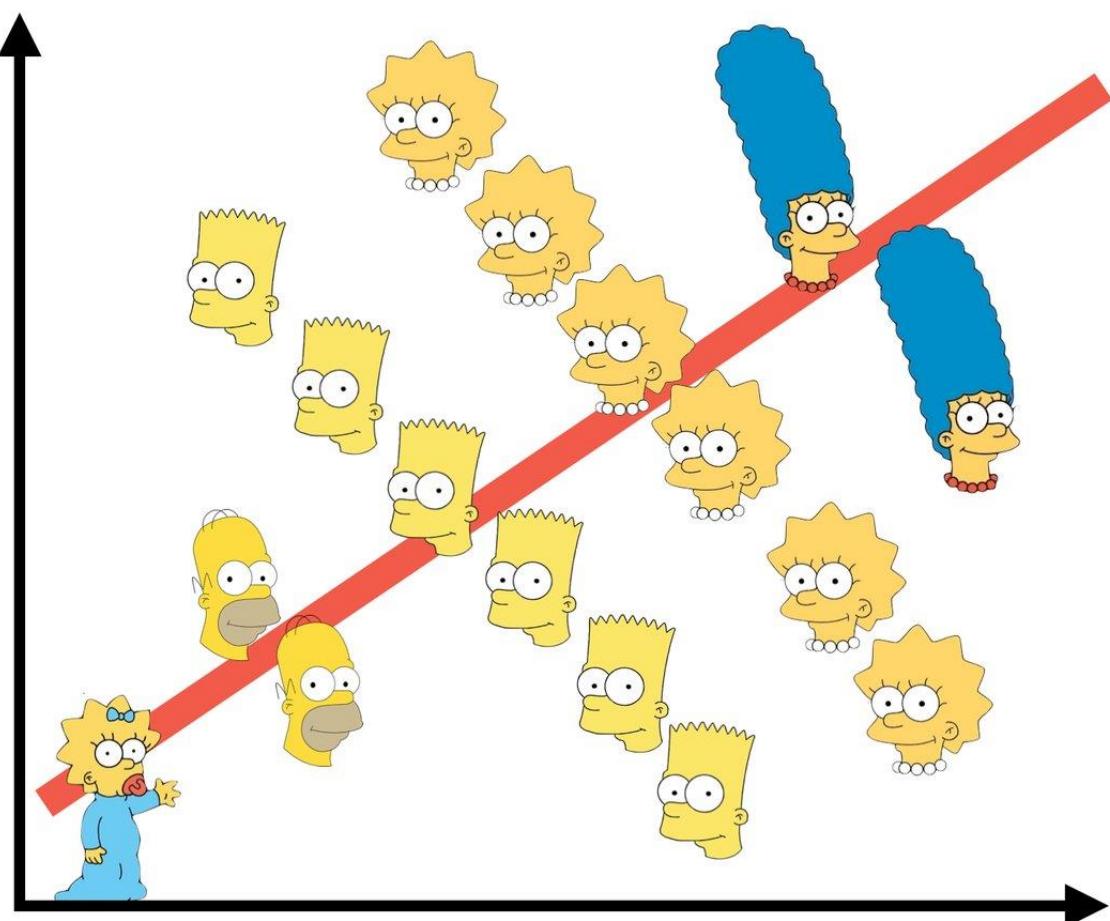
a. Dependent Variable: dag-salaris

- a) Hoeveel bedraagt de multipele correlatiecoëfficiënt? Interpreteer.
- b) Hoeveel bedraagt de determinatiecoëfficiënt? Interpreteer.
- c) Hoeveel respondenten zijn in de analyse betrokken?
- d) Hoeveel bedraagt de toetsstatistiek F ? Is deze significant ? Interpreteer.
- e) Hoeveel bedraagt het intercept en wat betekent dit?
- f) Hoeveel bedragen de ongestandaardiseerde richtingscoëfficiënten?
- g) Hoeveel bedragen de gestandaardiseerde richtingscoëfficiënten?
- h) Welke variabele heeft het sterkste relatieve effect? Welke parameter interpreteer je dat? Waarom?
- i) Zijn de regressieparameters significant?
- j) Wat betekent het 95% betrouwbaarheidsinterval voor B?
- k) Wat is het verwachte dag-salaris van een topmodel van 18jaar, zonder ervaring en met een gepercipieerde schoonheid van 81?

HOOFDSTUK 10

COMPLEXERE RELATIES TUSSEN VARIABELEN

SIMPSON'S PARADOX



HOOFDSTUK X

PADMODEL: INLEIDING

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk zijn studenten in staat uit een pad-diagram de bivariate correlaties te berekenen tussen een aantal variabelen. Studenten kunnen uit een padmodel de rechtstreekse, onrechtstreekse en totale effecten berekenen, interpreteren en rapporteren. Studenten kunnen bivariate correlaties tussen variabelen in een padmodel berekenen.

2. TE ONTHOUDEN KERNBEGRIPPEN

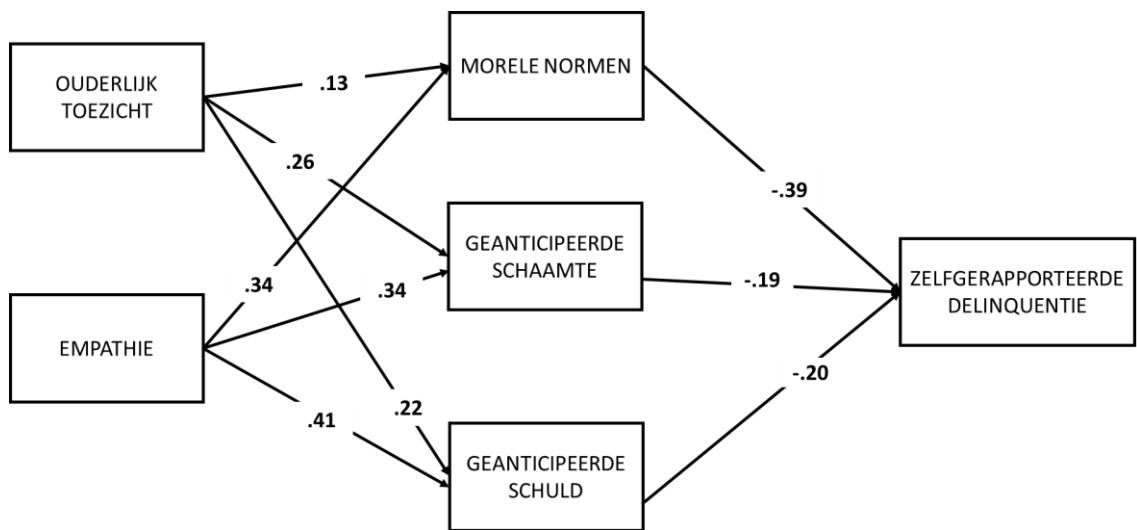
Conceptueel model of diagram	Concepten en hun onderlinge relaties
Endogene variabele	Afhankelijke variabele Variabele die causaal bepaald wordt door andere variabele(n)
Exogene variabele	Onafhankelijke variabele Variabele waar geen enkele causale effectrelaties toekomen
Mediator variabele of mediërende variabele	Variabele die tussen een oorzaak-gevolg relatie staat en het effect tussen oorzaak-gevolg beter verklaart en de relatie sterker maakt.
Onrechtstreeks of mediërend effect	Causaal pad walangs een variabele een effect uitoefent op een andere variabele maar nooit rechtstreeks, altijd via de impact op een tussenliggende variabele
Pad analyse	Een analyse van structurele modellen waarbij alle variabelen geobserveerd (manifest) zijn
Padmodel of paddiagram	Model van relaties tussen variabelen op basis van empirisch onderzoek
Spurieuze correlatie	In een pad analyse model gerepresenteerd door een gemeenschappelijke oorzaak te veronderstellen
Structureel model	Represents alle causale hypothese omvat de patronen van direct en indirecte effecten tussen alle variabelen in een statistisch model
Totale effect	Som van de rechtstreekse en onrechtstreekse effecten

3. OEFENINGEN

1. Empathie wordt vaak beschouwd als een belangrijke menselijke eigenschap. Een tekort aan empathische gevoelens wordt in verband gebracht met antisociaal gedrag en delinquentie (zie bijvoorbeeld Jolliffe & Farrington 2004, 2007)¹². Volgens Martin Hoffman³ (2001) ontwikkelt empathie zich al heel vroeg in het leven. Hoffman beschreef in zijn theoretisch model de ontwikkeling van empathie vanaf de geboorte tot volwassenheid waarbij het belang ervan aantoon voor de morele ontwikkeling en de morele emoties (vooral het ontstaan van schuldgevoelens). Empathie zou essentieel zijn voor een goede werking van ons geweten. Daarnaast is de rol van de ouders essentieel. Vooral de mate waarin ouders toezicht houden op het gedrag van hun kind en zijn/haar gedrag reguleren.

Hieronder vind je een criminologische toepassing van Hoffmans empathieontwikkelingsmodel.

We veronderstellen dat 'ouderlijk toezicht' en 'empathie' geen direct effect hebben op zelfgerapporteerde delinquentie maar dat het effect *indirect* is via drie dimensies van moraliteit: 'morele normen', 'geanticipeerde schaamte' en 'geanticipeerde schuld'. De variabelen werden gemeten aan de hand van gesommeerde likertschalen en mogen beschouwd worden als metrische variabelen.



¹ Jolliffe, D., & Farrington, D.P. (2004). Empathy and offending: A systematic review and meta-analysis. *Aggression and violent behavior*, 9(5), 441-476.

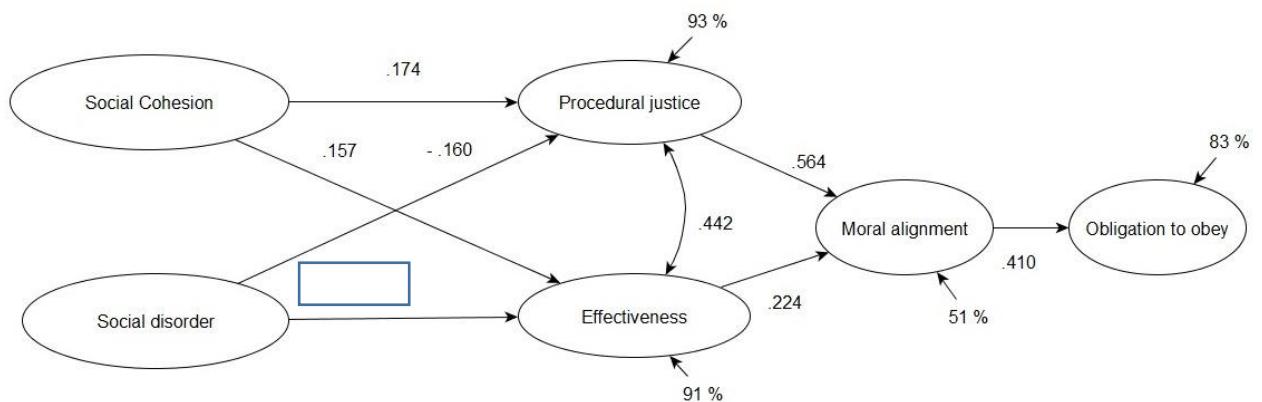
² Jolliffe, D., & Farrington, D.P. (2007). Examining the relationship between low empathy and self-reported offending. *Legal and Criminological Psychology*, 12(2), 265-286.

³ Hoffman, M.L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

- Welke variabelen in het model zijn 'exogene' variabelen ?
- Welke variabelen in het model zijn 'endogene' variabelen ?
- Welke variabelen in het model zijn 'intermediaire' variabelen ?
- Hoeveel bedraagt de correlatie tussen 'empathie' en 'zelfgerapporteerde delinquentie' ?
- Hoeveel 'indirecte' effecten van 'ouderlijk toezicht' op 'zelfgerapporteerde delinquentie' worden in het model weergegeven ?
- Hoeveel bedraagt het totale effect van 'ouderlijk toezicht' op 'zelfgerapporteerde delinquentie' ?

2. Onderstaand padmodel toont de resultaten van een uitbereiding op de toets van de procedurele rechtvaardigheidstheorie. Meer bepaald wordt er verondersteld dat percepties van sociale cohesie enerzijds en percepties van sociale desorganisatie anderzijds van invloed zijn op zowel het vertrouwen in de procedurele rechtvaardigheid van politie als het vertrouwen in de effectiviteit van politie. De achterliggende idee hier is dat burgers de politie verantwoordelijk achten voor o.m. de problemen in hun buurt. Verder wordt er verondersteld dat beide componenten van vertrouwen van invloed zijn op de mate waarin mensen het gevoel hebben dat politie dezelfde waarden en normen deelt als hen (moral alignment). Deze laatste variabele zou verder van invloed zijn op de mate waarin mensen zich verplicht voelen politie te gehoorzamen. De variabelen werden aan de hand van gesommeerde likertschalen gemeten waardoor je ze als ratiovariabelen mag beschouwen.

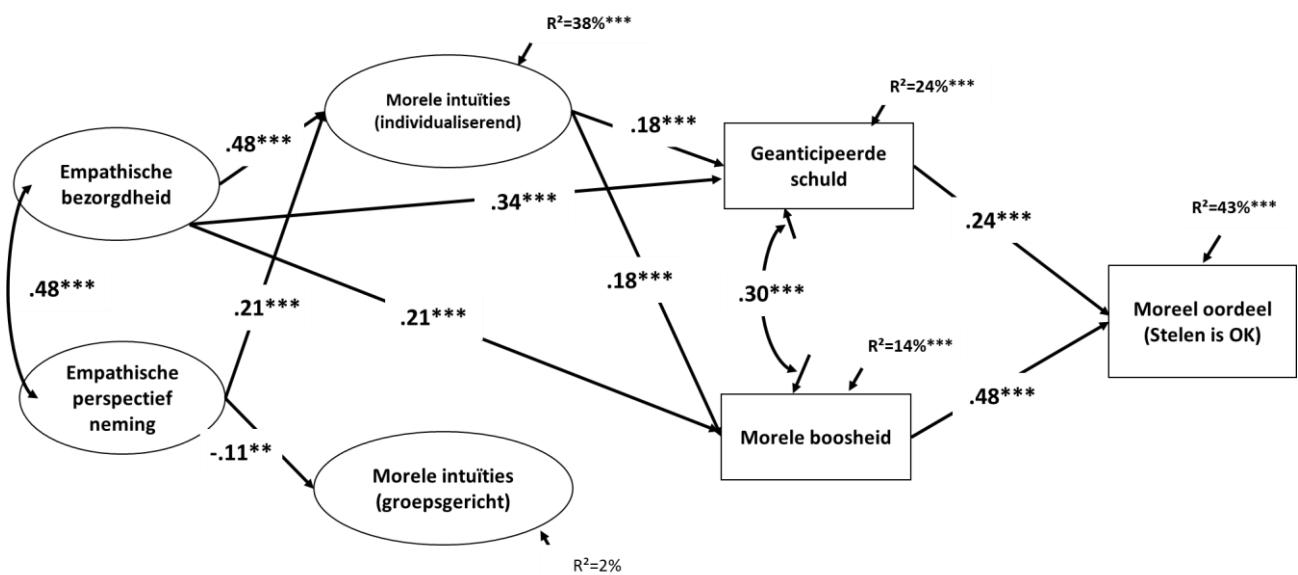
Het totale effect van Social disorder op Moral alignment bedraagt -0.1348.



- A) Hoeveel bedraagt het totale effect van social cohesion op obligation to obey?
- B) Hoeveel bedraagt het rechtstreekse effect van social disorder op effectiveness? Wat betekent dit?
- C) Hoeveel % van de variantie in obligation to obey kan verklaard worden door de variabelen in het model?
- D) Hoeveel bedraagt de onverklaarde variantie voor gepercipieerde effectiviteit van politie?

3. Het hieronder afgebeelde padmodel is een visualisatie van een partiële test van de Morele Fundamenten Theorie (MFT) van Jonathan Haidt en collega's⁴. MFT werd gecreëerd om te verklaren hoe en waarom moraliteit zo sterk varieert tussen culturen en tevens zoveel overeenkomsten en terugkerende thema's vertoont. De theorie vertrekt daarbij vanuit een evolutionair kader. De centrale vraag waarop de theorie een antwoord formuleert is *Wat zijn de determinanten van een moreel oordeel?* Volgens MFT zijn morele oordelen het resultaat van morele intuïties en morele emoties. In een recente studie (De Buck & Pauwels, 2022) werden de relaties onderzocht tussen dimensies van empathie en moreel oordelen in de context van verschillende vorm van normovertreding. In één geval werd een vignette gebruikt waarin een diefstal wordt beschreven. Aan respondenten werd gevraagd in welke mate zij stelen/diefstal als verkeerd beschouwen.

Je vindt hieronder het padmodel dat om didactische redenen is vereenvoudigd. Beantwoord onderstaande vragen.



⁴ Bron: De Buck, A. & Pauwels, L. J. R. (2022). *Explaining judgments on rule violations. On empathy, moral intuitions, and emotions*. Springer.

- Welke variabelen in het model zijn 'exogene' variabelen ?
- Welke variabelen in het model zijn 'endogene' variabelen ?
- Welke variabelen in het model zijn 'intermediaire' variabelen ?
- Hoeveel bedraagt de correlatie tussen 'morele intuïties (individueel)' en 'geanticipeerde schuld' ?
- Hoeveel 'indirecte' effecten van 'empathische perspectiefneming' op 'moreel oordeel' worden in het model weergegeven ?
- Hoeveel bedraagt het totale effect van 'empathische bezorgdheid' op 'geanticipeerde schuld' ?
- Hoeveel procent van de variatie in 'morele intuïties (groepsgericht)' kan verklaard worden op basis van de twee dimensies van empathie?

- Hoeveel procent van de variatie in de uitkomstvariabele kan niet verklaard worden op basis van dit model?

- Wat betekent dit concreet?

HOOFDSTUK 11

SYNTHÈSE OEFENING



HOOFDSTUK XI

SYNTHESE OEFENING

Hierna vind je een synthese oefening waarmee je zelf je kennis kan testen over de univariate, bivariate en meervoudige regressie analyse op basis van één databestand. Tracht onderstaande vragen te beantwoorden.

Dit is een voorbeeldoefening ter illustratie van de wijze waarop de examenvragen kunnen gesteld worden. Ter controle vind je de output van SPSS met de juiste coëfficiënten achteraan dit boek. Wees eerlijk tegenover jezelf en kijk niet eerst naar de oplossing. SPSS rond pas op het einde af en met 4 cijfers na de komma, dus er kan een minimale foutenmarge op je uitkomst zitten. Maak je daar geen zorgen over.

Veel succes met deze samenvattende oefening!

Op basis van een kleine toevalsstreekproef van twintig adolescenten werden de volgende data verzameld met betrekking tot biologisch geslacht, de mate van impulsiviteit (X_1), geanticipeerde schuld (X_2) en mate van moraliteit (mate waarin respondenten een aantal kleine delicten al dan niet goedkeuren) (Y).

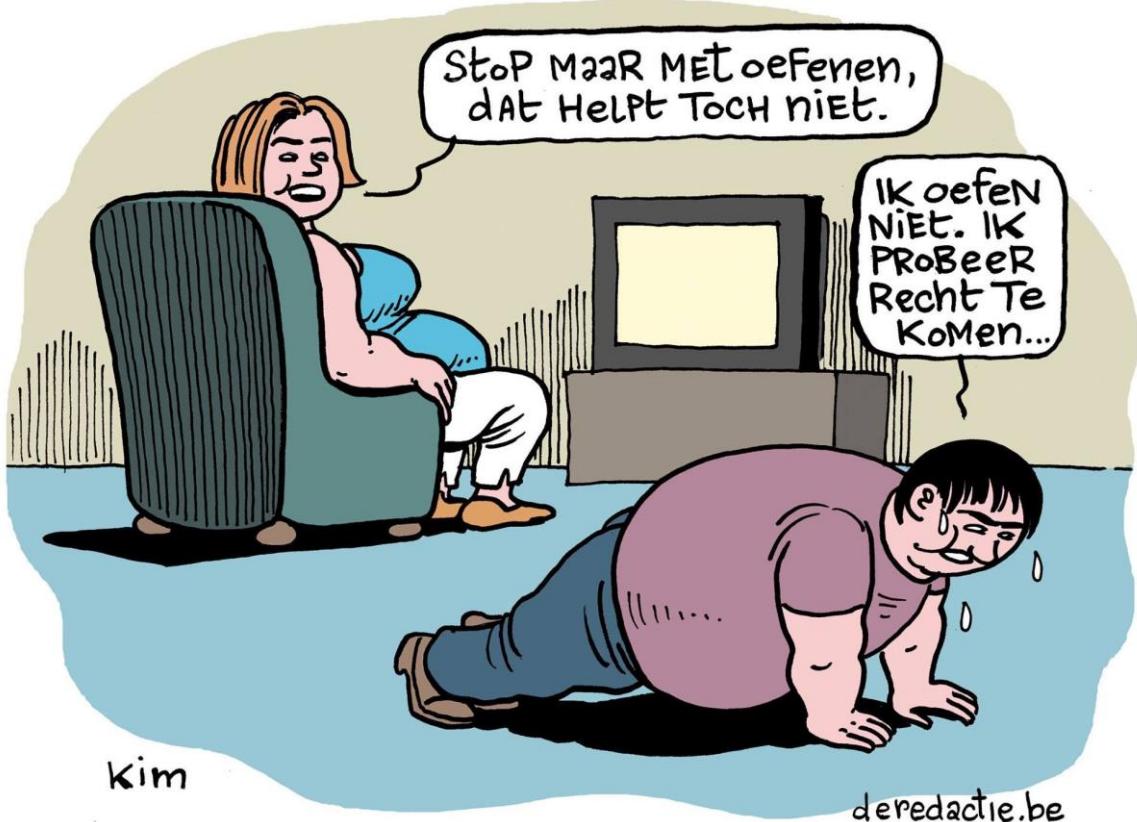
Adolescent	Geslacht Code 1 = jongen Code 2 = meisje	VARIABELE	VARIABELE	VARIABELE
		X1	X2	Y
		IMPULSIVITEIT	GEANTICIEERDE SCHULD	MORALITEIT
1	1	14	12	13
2	1	11	13	12
3	1	11	15	12
4	1	14	10	8
5	1	8	10	18
6	1	8	15	11
7	1	10	5	9
8	1	11	10	16
9	1	11	6	9
10	1	10	9	14
11	2	7	18	18
12	2	18	6	17
13	2	11	10	11
14	2	10	7	14
15	2	13	5	13
16	2	14	11	9
17	2	16	7	9
18	2	18	5	11
19	2	8	5	10
20	2	9	6	7

- 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 'moraliteit'(Y) kan verklaard worden op basis van 'impulsiviteit' (X1) en 'geanticipeerde schuld (X2)?
 - Welke factor (X1 of X2) heeft relatief het sterkste effect op 'moraliteit'?
 - Wat is de verwachte moraliteit voor een adolescent met een score van 15 op 'impulsiviteit' en een score van 12 op 'geanticipeerde schuld'?

HOOFDSTUK 12

HERHALINGSOEFENINGEN

TeGeN 2030 ZAL EuroPeAAh Te Dik zijn



HOOFDSTUK XII
HERHALINGSVRAGEN

VRAAG 1

Hoe groter de box in een boxplot, hoe meer waarnemingen die bevat.

- A. Deze uitspraak is fout
- B. Deze uitspraak is juist

VRAAG 2

In onderstaande tabel vind je de antwoorden op een vraag uit de Belgische editie van een internationaal zelfrapportage onderzoek dat peilt naar daderschap, slachtofferschap en middelengebruik onder adolescenten. Je ziet de frequentieverdeling van de antwoorden van Belgische respondenten op de vraag in welke mate de stelling '*mijn ouders weten waar ik ben als ik wegga*' op zichzelf van toepassing is. Respondenten konden kiezen uit 5 antwoordmogelijkheden: bijna altijd – vaak – soms – zelden – nooit.

Welke 2 uitspraken zijn juist? Uitspraken die je niet kan controleren aan de hand van de informatie die je hier krijgt, moet je als verkeerd beschouwen.

- A. De modus is nooit.
- B. 66,9% van de respondenten beantwoordde de stelling '*mijn ouders weten waar ik ben als ik wegga*' met zelden of nooit.
- C. 33,1 % van de respondenten beantwoordde de stelling '*mijn ouders weten waar ik ben als ik wegga*' met soms.
- D. De waarde voor het eerste kwartiel bedraagt 'vaak'.
- E. 4673 respondenten hebben een geldig antwoord gegeven op de stelling '*mijn ouders weten waar ik ben als ik wegga*'.

<i>Mijn ouders weten waar ik ben als ik wegga</i>					
		Frequentie	Percentage	Geldig percentage	Cumulatief percentage
Geldig	Bijna altijd	479	10,1	10,3	10,3
	Vaak	360	7,6	7,7	18,0
	Soms	705	14,8	15,1	33,1
	zelden	1114	23,4	23,9	57,0
	nooit	2005	42,1	43,0	100,0
	Total	4663	98,0	100,0	
Missing	Ambigue antwoord	10	,2		
	Geen antwoord	85	1,8		
	Totaal	95	2,0		
Totaal		4758	100,0		

VRAAG 3

Een masterstudent voert een verkennend onderzoek uit naar factoren die een effect kunnen hebben op zelfgerapporteerde regelovertreding. Op basis van een grondige literatuurstudie identificeert de student een aantal variabelen die van belang zijn, met name: 'ouderlijk toezicht', 'schoolcriminaliteit', 'deviante vrienden', 'antisociale normen en waarden', 'lage capaciteit tot zelfcontrole' en 'anticipatie op schaamtegevoelens'; Hierna vinden jullie de SPSS output van de multiple regressieanalyse die de student heeft uitgevoerd. Een aantal waarden werden bewust weggelaten uit de output.

Welke twee uitspraken zijn juist:

- A. de nulhypothese dat schoolcriminaliteit geen effect heeft op zelfgerapporteerde regelovertreding kunnen we niet verwerpen
op p: ≤0.05 niveau.
- B. Een respondent met scores 0 op ouderlijke toezicht, schoolcriminaliteit en deviante vrienden, met score 2 op antisociale normen, met score 1.5 op lage capaciteit tot zelfcontrole en met score -1 op anticipatie op schaamtegevoelens, zou theoretisch gezien een score van 0.797 hebben voor zelfgerapporteerde criminaliteit.
- C. Op basis van deze analyse kan 66.7% van de variabiliteit in zelfgerapporteerde regelovertreding verklaard worden door alle onafhankelijke variabelen opgenomen in de analyse.
- D. De kritieke F-waarde die correspondeert met deze analyse bedraagt 470,578.
- E. Er zijn 3536 respondenten in de analyse betrokken.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	,667 ^a	,444	,443	,75235865	,444	470,539	6	3530	,000

a. Predictors: (Constant), anticipatie op schaamtegevoelens, schoolcriminaliteit, ouderlijk toezicht, devante vrienden, antisociale normen en waarden, lage capaciteit tot zelfcontrole

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1598,074				
	Residual	1998,134	3530			
	Total	3596,208	3536			

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Coefficients	Beta		
1	(Constant)	,091	,013			7,122	,000
	Ouderlijk toezicht	-,042	,014		-,040	-2,978	,003
	schoolcriminaliteit	,026	,014		,024	1,877	,061
	Devante vrienden	,459	,015		,426	29,842	,000
	Antisociale normen en waarden	,192	,015		,186	12,692	,000
	Lage capaciteit tot zelfcontrole	,124	,014		,120	8,554	,000
	Anticipatie op schaamtegevoelens	-,136	,015		-,128	-8,952	,000

a. Afhankelijke variabele: zelfgerapporteerde regelovertrreding

VRAAG 4

In onderstaande tabel worden de ruwe prestatie-scores weergegeven van 12 atleten. We willen statistisch nagaan of de trainingsgroep die 7u of meer traint significant verschilt van de trainingsgroep die 6u of minder traint voor wat betreft de prestatiescores.

Welke twee uitspraken zijn correct? Rond telkens af op twee decimalen.

- A. 72,9% van de variatie in de waargenomen waarden op prestatiescores kan verklaard worden door de trainingsgroepen.
- B. De totale tussengroepsvariantie bedraagt 1704,08.
- C. De corresponderende F-waarde in deze analyse bedraagt 21,04 en is significant.
- D. De totale binnengroepsvariantie bedraagt 632,83.
- E. De kritieke F-waarde gegeven het aantal vrijheidsgraden in teller en in noemer bij een significantieniveau van $p \leq 0.001$ bedraagt 26,92.

Observaties	Gender	Prestatie-scores op 100 (met 100= excellent)	Uren training
1	Meisje	67	6u of minder
2	Jongen	87	7u of meer
3	Meisje	87	7u of meer
4	Jongen	56	6u of minder
5	Meisje	72	7u of meer
6	Jongen	57	6u of minder
7	Meisje	60	6u of minder
8	Jongen	92	7u of meer
9	Meisje	56	6u of minder
10	jongen	67	7u of meer
11	Meisje	56	6u of minder
12	Jongen	90	7u of meer

VRAAG 5

We presenteren hieronder de correlatiematrix met de bivariate correlaties tussen 6 variabelen afkomstig uit verschillende criminologische theorieën. De variabelen zijn geoperationaliseerd aan de hand van gesommeerde likertschalen en kunnen als ratiovariabelen beschouwd worden.

Welke twee uitspraken zijn juist ?

- A. We kunnen zeggen dat niveaus van 'schuldgevoelens' zullen stijgen met 16,65 wanneer 'band met ouders' met 1 eenheid stijgt.
- B. Wanneer iemand 0 scoort op de schaal 'band met de ouders' en 0 op de schaal die peilt naar 'delinquentietolerantie' wordt verwacht dat hij/zij 24,393 zal scoren op de schaal die peilt naar 'schuldgevoelens'.
- C. Wanneer we het bivariate verband bekijken tussen 'band met ouders' en 'delinquentietolerantie' onder de controle van 'schuldgevoelens' kunnen we vaststellen dat het bivariate verband zwakker is geworden, met name een daling van -.324 naar -.155.
- D. De bivariate analyse verklaart 52,5 % van de variabiliteit tussen 'delinquentietolerantie' en 'deviate vrienden'
- E. 'Delinquentietolerantie' blijkt een betere voorspeller te zijn voor het ervaren van 'schuldgevoelens' dan 'band met ouders'.
- F. 'Band met ouders' en 'delinquentietolerantie' verklaren samen 65,7% van de variabiliteit in geobserveerde waarden voor 'schuldgevoelens'

CORRELATIEMATRIX

	1	2	3	4	5	6
1. Band met ouders	1					
2. Band school	,340	1				
3. Delinquentietolerantie	-,324	-,394	1			
4. Schaamtegevoelens	.297	.274	-.522	1		
5. Schuldgevoelens	.329	.358	-.645	.720	1	
6. Deviate vrienden	-,326	-,281	.525	-,475	-,514	1

Alle correlaties significant op het niveau p:<.01

N= 843

Beschrijvende statistieken

	N	Minimum	Maximum	Mean	Std. Deviation
Band ouders	843	9,00	28,00	22,8309	3,44260
Band school	843	13	33	25,73	3,396
delinquentietolerantie	843	16	64	33,83	9,492
Schaamtegevoelens	843	6	18	14,89	3,124
Schuldgevoelens	843	6	18	13,00	2,909
Devante vrienden	843	6	23	9,04	3,444
Valid N (listwise)	843				

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			Sig. F Change
						F Change	df1	df2	
1	,657 ^a	,432	,431	2,195	,432	319,341	2	840	,000

a. Predictors: (Constant), delinquentietolerantie, band met ouders

Coefficients^a

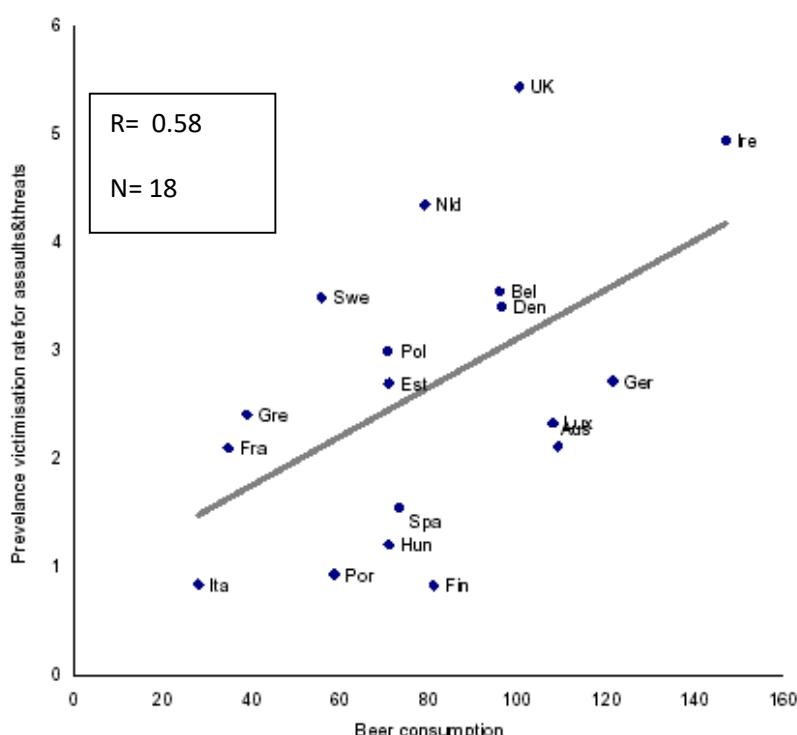
Model		Unstandardized Coefficients		Standardized Coefficients		Correlations	
		B	Std. Error	Beta	t		
		(Constant)					
1	Band met ouders	,113	,023	,134	4,873	,000	,329
	delinquentietolerantie	-,184	,008	-,601	-21,881	,000	-,645
							-,603
							-,569

a. afhankelijke variabele: schuldgevoelens

VRAAG 6

De onderstaande puntenwolk geeft de relatie weer tussen de nationale prevalentie slachtofferschap geweld/dreigen met geweld en de nationale bierconsumptie per 100.000 inwoners (N= 18). Beoordeel de uitspraken.

Figure 2.21 National prevalence of Assaults/threats and consumption of beer per 100,000 of the population.



Uitspraken	Waar	Vals
Hoe hoger de bierconsumptie per land, hoe hoger het slachtofferpercentage van geweldsdelen.		
Bierdrinkers zijn de plegers van het geweld.		
Als we België zouden weglaten, zou de regressielijn er compleet anders uit zien.		
Het Verenigd Koninkrijk is een outlier in deze analyse.		
De determinatiecoëfficiënt bedraagt 33.64%		
In Duitsland is het slachtofferpercentage lager dan wat men zou verwachten op basis van de bierconsumptie van de Duitsers.		
Het residu van Denemarken is negatief.		

VRAAG 7

Een bekend consumentenonderzoeksbureau doet een vergelijkend onderzoek naar de kwaliteit van koffiepads van verschillende merken. Ze geven per fabrikant een score op 10 voor de smaak van hun koffie en geven daarnaast ook weer hoeveel percent van de gemalen koffie in de pads van Braziliaanse afkomst is.

Hoeveel punten verwacht je dat een nieuwe fabrikant zal krijgen als je weet dat zijn koffiepads 40% Braziliaanse koffie bevatten ? (afgerond naar het dichtstbijzijnde gehele getal)

Duid één correct antwoord aan.

Fabrikant	Braziliaanse afkomst (in %)	Smaak (score op 10)
1	36	6
2	18	3
3	11	4
4	24	5
5	45	7
6	63	8
7	4	4
8	13	4
9	74	8
10	4	3

- A. Zes op tien
- B. Zeven op tien
- C. Vijf op tien
- D. Acht op tien

VRAAG 8

De meervoudige regressieanalyse met X1 en X2 als onafhankelijke variabelen geeft hetzelfde resultaat als twee afzonderlijke bivariate regressieanalyses. Is dit een juiste uitspraak ?

- Deze uitspraak is enkel juist in de situatie waarbij $r(x_1, x_2) = 0$.
- Deze uitspraak is juist.
- Deze uitspraak is steeds fout.

VRAAG 9

De organisatie die waakt over de auteursrechten op onder andere muziek, SABAM, wenst te onderzoeken of ook particulieren moeten betalen voor de muziek die ze thuis laten afspelen. Dit kadert in de strijd tegen piraterij op het internet. Om te monitoren wie, wat en hoeveel mensen muziek laat spelen thuis, voert de organisatie een pilootonderzoek uit bij 1000 Vlamingen. Een opmerkelijk resultaat situeert zich tussen de verschillende geslachten met betrekking tot de antwoorden op de vraag: *“Hoe vaak speelt u thuis muziek af per week?”*

Duid het correcte antwoord aan. (1 JUIST antwoord)

	< 10	> 10
Mannen	153	296
Vrouwen	221	330

Bereken Cramer's V.

- A. Cramer's V bedraagt 0.06. De relatie tussen beide variabelen is uitermate zwak.
- B. Cramer's V bedraagt 0.16. De relatie tussen beide variabelen is zwak.
- C. Cramer's V bedraagt 0.25. De relatie tussen beide variabelen is zwak tot matig.
- D. Cramer's V bedraagt 0.33. De relatie tussen beide variabelen is matig.

VRAAG 10

Heteroscedasticiteit wil zeggen dat...

- De waarde van de residuele termen toenemen naarmate X_1 toeneemt.
- De waarde van de residuele termen gelijk blijft, naarmate X_1 toeneemt.

VRAAG 11

We deden onderzoek naar lidmaatschap van problematische jeugdgroepen in België. We vroegen aan 4408 Belgische jongens en meisjes of zij al dan niet lid waren van een problematische jeugdgroep. Hieronder vind je de kruistabel met de gerapporteerde antwoorden. Welke twee uitspraken zijn juist ?

Bij de berekening van Chi² rond je in de tussenstappen af op 4 decimalen, in de uitkomst tot op 2 decimalen.

- A. Bij jongens ligt de verhouding tussen wel lid en geen lid van problematische jeugdgroepen in België 0,14 keer hoger dan bij meisjes.
- B. Het relevante percentageverschil bedraagt 2,6 percentagepunten.
- C. Jongens hebben 7,08 keer meer kans om betrokken te zijn bij een problematische jeugdgroep dan meisjes.
- D. Chi² bedraagt 7,34
- E. Het relevante percentageverschil bedraagt 12,4%

	Lid van problematische jeugdgroep		Totaal
	GEEN LID	WEL LID	
MEISJES	2006	218	2224
JONGENS	1914	270	2184
	3920	488	4408

VRAAG 12

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.

VRAAG 13

Interactie betekent dat het effect van X1 op Y conditioneel is op X2

- Deze uitspraak is juist.
- Deze uitspraak is verkeerd.

VRAAG 14

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

VRAAG 15

De standaardafwijking is een robuuste parameter van centraliteit

- Deze uitspraak is juist.
- Deze uitspraak is verkeerd.

VRAAG 16

Irina onderzoekt voor haar masterproef het effect van discriminatie op het mentaal welbevinden van migranten. Beide worden gemeten op een schaal van 0 tot 10. Een hogere score op de onafhankelijke variabele betekent het vaker ervaren van discriminatie; terwijl een hogere score op afhankelijke variabele staat voor een betere mentale gezondheid. Ze vindt een covariantie van -0,248; een standaardafwijking van 1,6 voor x en een variantie van 0,0625 voor y. Welke stelling is JUIST? (1 antwoord is correct).

- a. 62 % van de variantie in het mentaal welbevinden van migranten wordt verklaard door de variantie van de mate waarin migranten discriminatie ondervinden.
- b. 38,44 % van de variantie in het mentaal welbevinden van migranten wordt verklaard door de variantie van de mate waarin ze gediscrimineerd worden.
- c. Bij een verschil van 1 standaardafwijking op de schaal over het mentaal welbevinden van migranten verwachten we 0,62 standaardafwijkingen verschil op de schaal rond discriminatie die migranten ondervinden.
- d. Bij een verschil van 1 op de schaal van discriminatie, verwachten we een verschil van 0,38 op de schaal die het mentaal welbevinden van migranten meet.

VRAAG 17

Een criminologe is geïnteresseerd in de effecten van anomie (normenloosheid) op het plegen van zelfmoord. Op basis van een aantal sociaaleconomische parameters verdeelt zij 15 steden in 3 groepen : hoge anomie - matige anomie en lage anomie. Vervolgens bekomt zij de zelfmoordgraad voor elke groep (aantal zelfmoorden per 100.000 inwoners).

Welke statistische procedure zou jij toepassen om te bepalen of er een statistisch significant verschil is in zelfmoordgraad naargelang het niveau van anomie (normenloosheid) ?

Duid één juist antwoord aan.

- A. Berekenen van Chi² en Cramer's V.
- B. Meervoudige regressie analyse.
- C. One-way variantie analyse.
- D. Berekenen van Pearson correlatiecoëfficiënt.

VRAAG 18

Een bedrijf wenst iemand aan te werven. Tijdens de eerste selectieronde moeten de sollicitanten ($n=803$) een psychotechnische proef afleggen. De resultaten zijn perfect normaal verdeeld $N(45,8)$. Alleen de sollicitanten met een score hoger dan 59 worden uitgenodigd voor een gesprek. Hoeveel sollicitanten zullen dit zijn (afgerond cijfer)?

- A. Afgerond 771 sollicitanten
- B. 0 sollicitanten
- C. Afgerond 612 sollicitanten
- D. Afgerond 51 sollicitanten
- E. Afgerond 32 sollicitanten

VRAAG 19

Het significantieniveau α is de kans dat een significantietoets de nulhypothese (H_0) zal verwerpen terwijl H_0 in feite juist is.

Deze uitspraak is

- A. JUIST
- B. FOUT

VRAAG 20

Het aantal inbraken per jaar in een bepaald arrondissement is normaal verdeeld: $N(311.50)$.

Welke uitspraak is juist ?

- A. De kans dat in 2017 minder dan 250 inbraken worden geregistreerd is 11%.
- B. De kans dat in 2017 meer dan 350 inbraken geteld worden, is 78.23%.
- C. De kans dat tussen de 300 en 350 inbraken worden genoteerd, is 12%.
- D. De kans dat tussen de 350 en 375 inbraken worden genoteerd, is 37%.

BIBLIOGRAFIE

- Dancey, C., & Reidy, J. (2017). *Statistics without maths for psychology. Seventh Edition.* Pearson.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics.* 5th Edition: SAGE edge.
- Fox, J. A., Levin, J., & Shively, M. (2002). *Elementary statistics in criminal justice research.* Pearson Education.
- Howitt, D., & Cramer, D. (2007). *Statistiek in de sociale wetenschappen, 3/e.* Pearson Education.
- Huizingh E., (2014). *Inleiding SPSS 22 voor IBM SPSS Statistics.* Twaalfde editie. Academic Service, Sdu Uitgevers bv. BIM Media B.V., Den Haag.
- McClave, J. T., & Sincich, T., & Knypstra, S. (2016). *Statistiek, 12/e.* Pearson Education.
- Moore, D. S., & McCabe, G. P. (2005). *Statistiek in de praktijk: theorieboek.* Sdu Uitgevers bv, Den Haag.
- Moore, D.S., & McCabe, G.P. (2005) *Statistiek in de praktijk : Opgavenboek.* Sdu Uitgevers bv, Den Haag, Academic Service.
- Moore, D.S., McCabe, G.P., & Craig, B.A. (2021). *Introduction to the practice of statistics. Tenth Edition.* MacMillan International Higher Education.
- Mortelmans, D., & Dehertogh, B. (2006). *Kennismaken met SPSS en SAS.* Acco: Leuven/Voorburg.
- Pauwels, L. (2012). *Toegepaste Statistiek met SPSS voor criminologen.* Maklu, Antwerpen-Apeldoorn.
- te Grotenhuis, M., & van der Weegen, T. (2008). *Statistiek als hulpmiddel. Een overzicht van gangbare toepassingen in de sociale wetenschappen.* Van Gorcum.
- Weisburd, D., & Britt, C. (2007). *Statistics in criminal justice. Third Edition.* Springer.
- Weisburd, D., Wilson, D.B., Wooditch, A., & Britt C. (2022). *Advanced statistics in criminology and criminal justice. Fifth Edition.* Springer.

OPLOSSING

SYNTHESOEFENING

Hierna vind je de SPSS output van de extra synthese oefening (zie DEEL X).

Ter herinnering: SPSS rondt pas op het einde af en met 4 cijfers na de komma, dus er kan een minimale foutenmarge op je uitkomst zitten. Maak je daar geen zorgen over.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
IMPULSIVITEIT_X1	20	7,00	18,00	11,6000	3,21837	10,358
SCHULD_X2	20	5,00	18,00	9,2500	3,86448	14,934
MORALITEIT_Y	20	7,00	18,00	12,0500	3,30032	10,892
Valid N (listwise)	20					

Correlations

		IMPULSIVITEIT_X1	SCHULD_X2	MORALITEIT_Y
IMPULSIVITEIT_X1	Pearson Correlation	1	-,360	-,137
	Sig. (2-tailed)		,119	,565
	Sum of Squares and Cross-products	196,800	-85,000	-27,600
	Covariance	10,358	-4,474	-1,453
	N	20	20	20
SCHULD_X2	Pearson Correlation	-,360	1	,333
	Sig. (2-tailed)	,119		,151
	Sum of Squares and Cross-products	-85,000	283,750	80,750
	Covariance	-4,474	14,934	4,250
	N	20	20	20
MORALITEIT_Y	Pearson Correlation	-,137	,333	1
	Sig. (2-tailed)	,565	,151	
	Sum of Squares and Cross-products	-27,600	80,750	206,950
	Covariance	-1,453	4,250	10,892
	N	20	20	20

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean			
						Lower Bound	Upper Bound	Minimum	Maximum
IMPULSIVITEIT_X1	1 JONGEN	10	10,8000	2,04396	,64636	9,3378	12,2622	8,00	14,00
	2 MEISJE	10	12,4000	4,03320	1,27541	9,5148	15,2852	7,00	18,00
	Total	20	11,6000	3,21837	,71965	10,0938	13,1062	7,00	18,00
SCHULD_X2	1 JONGEN	10	10,5000	3,37474	1,06719	8,0859	12,9141	5,00	15,00
	2 MEISJE	10	8,0000	4,08248	1,29099	5,0796	10,9204	5,00	18,00
	Total	20	9,2500	3,86448	,86412	7,4414	11,0586	5,00	18,00
MORALITEIT_Y	1 JONGEN	10	12,2000	3,19026	1,00885	9,9178	14,4822	8,00	18,00
	2 MEISJE	10	11,9000	3,57305	1,12990	9,3440	14,4560	7,00	18,00
	Total	20	12,0500	3,30032	,73797	10,5054	13,5946	7,00	18,00

Correlations

			IMPULSIVITEIT _X1	SCHULD_X2	MORALITEIT_Y
GESLACHT					
1 JONGEN	IMPULSIVITEIT_X1	Pearson Correlation	1	-,032	-,385
		Sig. (2-tailed)		,930	,272
		Sum of Squares and Cross-products	37,600	-2,000	-22,600
		Covariance	4,178	-,222	-2,511
		N	10	10	10
	SCHULD_X2	Pearson Correlation	-,032	1	,217
		Sig. (2-tailed)	,930		,548
		Sum of Squares and Cross-products		-2,000	102,500
		Covariance		-,222	11,389
		N	10	10	10
2 MEISJE	IMPULSIVITEIT_X1	Pearson Correlation	-,385	,217	1
		Sig. (2-tailed)	,272	,548	
		Sum of Squares and Cross-products		-22,600	21,000
		Covariance		-2,511	2,333
		N	10	10	10
	SCHULD_X2	Pearson Correlation	-,425	-,425	-,020
		Sig. (2-tailed)	,221	,221	,956
		Sum of Squares and Cross-products	146,400	-63,000	-2,600
		Covariance	16,267	-7,000	-,289
		N	10	10	10
	MORALITEIT_Y	Pearson Correlation	-,425	1	,427
		Sig. (2-tailed)	,221		,219
		Sum of Squares and Cross-products		-63,000	150,000
		Covariance		-7,000	16,667
		N	10	10	10

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
IMPULSIVITEIT_X1	Between Groups	12,800	1	12,800	1,252	,278
	Within Groups	184,000	18	10,222		
	Total	196,800	19			
SCHULD_X2	Between Groups	31,250	1	31,250	2,228	,153
	Within Groups	252,500	18	14,028		
	Total	283,750	19			
MORALITEIT_Y	Between Groups	,450	1	,450	,039	,845
	Within Groups	206,500	18	11,472		
	Total	206,950	19			

ANOVA Effect Sizes^{a,b}

		Point Estimate	95% Confidence Interval	
			Lower	Upper
IMPULSIVITEIT_X1	Eta-squared	,065	,000	,324
	Epsilon-squared	,013	-,056	,287
	Omega-squared Fixed-effect	,012	-,053	,276
	Omega-squared Random-effect	,012	-,053	,276
SCHULD_X2	Eta-squared	,110	,000	,379
	Epsilon-squared	,061	-,056	,344
	Omega-squared Fixed-effect	,058	-,053	,333
	Omega-squared Random-effect	,058	-,053	,333
MORALITEIT_Y	Eta-squared	,002	,000	,156
	Epsilon-squared	-,053	-,056	,109
	Omega-squared Fixed-effect	-,050	-,053	,104
	Omega-squared Random-effect	-,050	-,053	,104

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

Regressie van Y op X1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,137 ^a	,019	-,036	3,35890

a. Predictors: (Constant), IMPULSIVITEIT_X1

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,871	1	3,871	,343	,565 ^b
	Residual	203,079	18	11,282		
	Total	206,950	19			

a. Dependent Variable: MORALITEIT_Y

b. Predictors: (Constant), IMPULSIVITEIT_X1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13,677	2,877		4,754	<.001
	IMPULSIVITEIT_X1	-,140	,239	-,137	-,586	,565

a. Dependent Variable: MORALITEIT_Y

Regressie van Y op X2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,333 ^a	,111	,062	3,19696

a. Predictors: (Constant), SCHULD_X2

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22,980	1	22,980	2,248	,151 ^b
	Residual	183,970	18	10,221		
	Total	206,950	19			

a. Dependent Variable: MORALITEIT_Y

b. Predictors: (Constant), SCHULD_X2

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.
		B	Std. Error			
1	(Constant)	9,418	1,896		4,968	<.001
	SCHULD_X2	,285	,190	,333	1,499	,151

a. Dependent Variable: MORALITEIT_Y

Regressie van X1 op Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,137 ^a	,019	-,036	3,27549

a. Predictors: (Constant), MORALITEIT_Y

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,681	1	3,681	,343	,565 ^b
	Residual	193,119	18	10,729		
	Total	196,800	19			

a. Dependent Variable: IMPULSIVITEIT_X1

b. Predictors: (Constant), MORALITEIT_Y

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta			
1	(Constant)	13,207	2,840			4,651	<.001
	MORALITEIT_Y	-,133	,228	-,137		-,586	,565

a. Dependent Variable: IMPULSIVITEIT_X1

Regressie van X2 op Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,333 ^a	,111	,062	3,74345

a. Predictors: (Constant), MORALITEIT_Y

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31,508	1	31,508	2,248	,151 ^b
	Residual	252,242	18	14,013		
	Total	283,750	19			

a. Dependent Variable: SCHULD_X2

b. Predictors: (Constant), MORALITEIT_Y

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta			
1	(Constant)	4,548	3,245			1,401	,178
	MORALITEIT_Y	,390	,260	,333		1,499	,151

a. Dependent Variable: SCHULD_X2

Regressie van Y op X1 en X2

Descriptive Statistics

	Mean	Std. Deviation	N
MORALITEIT_Y	12,0500	3,30032	20
IMPULSIVITEIT_X1	11,6000	3,21837	20
SCHULD_X2	9,2500	3,86448	20

Correlations

	MORALITEIT_Y	IMPULSIVITEIT_X1	SCHULD_X2
Pearson Correlation	MORALITEIT_Y	1,000	-,137
	IMPULSIVITEIT_X1	-,137	1,000
	SCHULD_X2	,333	-,360
Sig. (1-tailed)	MORALITEIT_Y	,	,076
	IMPULSIVITEIT_X1	,283	,
	SCHULD_X2	,076	,060
N	MORALITEIT_Y	20	20
	IMPULSIVITEIT_X1	20	20
	SCHULD_X2	20	20

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	,334 ^a	,111	,007	3,28904	,111	1,065	2	17	,367

a. Predictors: (Constant), SCHULD_X2, IMPULSIVITEIT_X1

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23,048	2	11,524	1,065	,367 ^b
	Residual	183,902	17	10,818		
	Total	206,950	19			

a. Dependent Variable: MORALITEIT_Y

b. Predictors: (Constant), SCHULD_X2, IMPULSIVITEIT_X1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.
		B	Std. Error			
1	(Constant)	9,704	4,104		2,364	,030
	IMPULSIVITEIT_X1	-,020	,251	-,019	-,079	,938
	SCHULD_X2	,279	,209	,326	1,331	,201

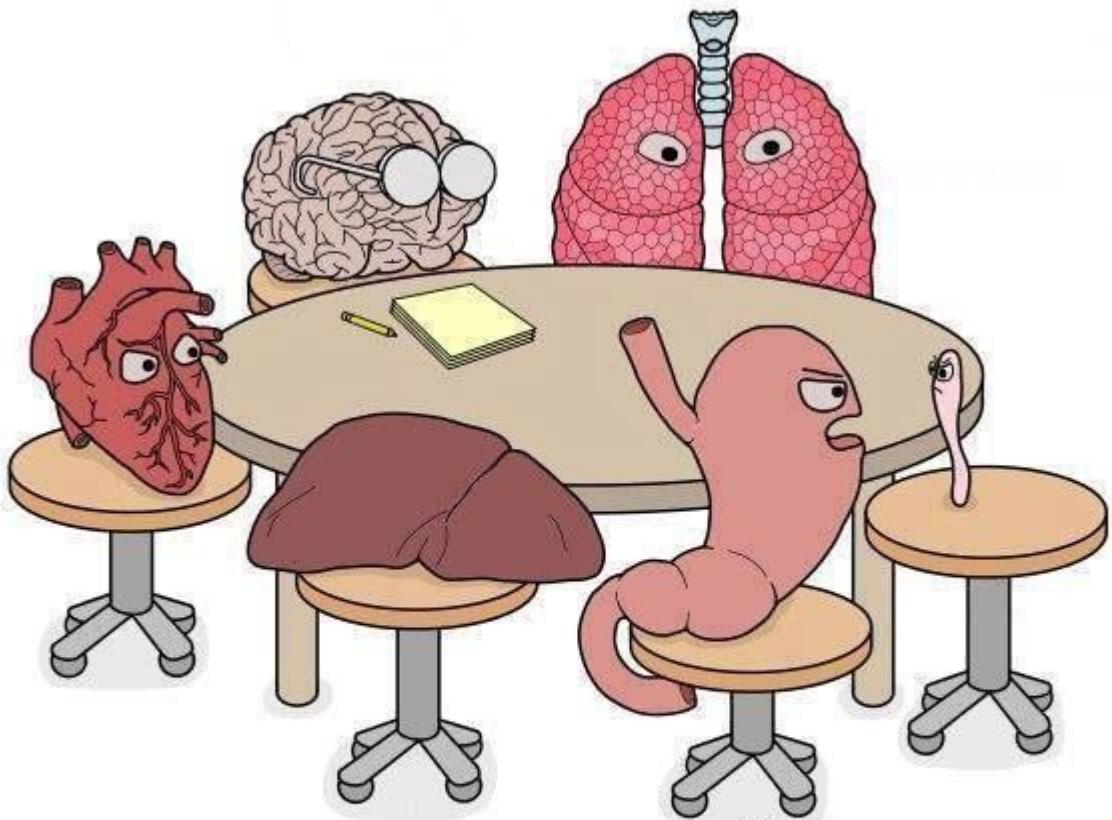
a. Dependent Variable: MORALITEIT_Y

Bereken de partiële correlatiecoëfficiënt tussen X1 en Y onder controle van X2.

Correlations

			IMPULSIVITEIT _X1	MORALITEIT_Y
Control Variables				
SCHULD_X2	IMPULSIVITEIT_X1	Correlation	1,000	-,019
		Significance (2-tailed)	,	,938
		df	0	17
	MORALITEIT_Y	Correlation	-,019	1,000
		Significance (2-tailed)	,938	,
		df	17	0

APPENDIX



**"Why do you keep coming to these meetings,
appendix? You have not contributed a
damn thing since the Paleolithic!"**

APPENDIX
BASISHANDLEIDING BIJ SPSS

1. DOELSTELLINGEN

Deze appendix is een eerste inleiding tot het statistische softwareprogramma SPSS en is bedoeld voor eerstejaarsstudenten Criminologische Wetenschappen. De nadruk ligt op het stap voor stap aanleren van de basisprincipes voor het efficiënt en verantwoord uitvoeren van eenvoudige statistische analyses. Deze basishandleiding SPSS geeft geïnteresseerde studenten de mogelijkheid op zelfstandige basis reeds kennis te maken met de meest recente versie van SPSS Statistics 25, de opbouw, werkwijze en mogelijkheden. Na het doorlopen van de instructies zijn studenten instaat om een eigen gegevensbestand op te maken en enkele eenvoudige analyses uit te voeren. In dit hoofdstuk worden ook voorbeelden gegeven hoe SPSS-output dient geïnterpreteerd en gerapporteerd. Echter, een grondiger inleiding volgt later tijdens de opleiding Criminologische Wetenschappen in het opleidingsonderdeel ‘Kwantitatieve Toegepaste data analyse en rapportage’.

2. WAT IS SPSS ?

SPSS staat voor *Statistical Package for the Social Sciences* en is een software programma om statistische analyses uit te voeren dat op maat werd geschreven voor gebruik in de sociale wetenschappen. Omwille van het succes ervan in andere wetenschappelijke disciplines en in zowel commerciële als niet-commerciële organisaties wordt nu gewoon van SPSS gesproken. Het is een zeer krachtig programma dat met een hoge snelheid uitgebreide datasets kan analyseren. Op een paar seconden tijd kan SPSS het werk verrichten waarvoor een onderzoeker een tiental jaren terug nog meerdere dagen nodig had.

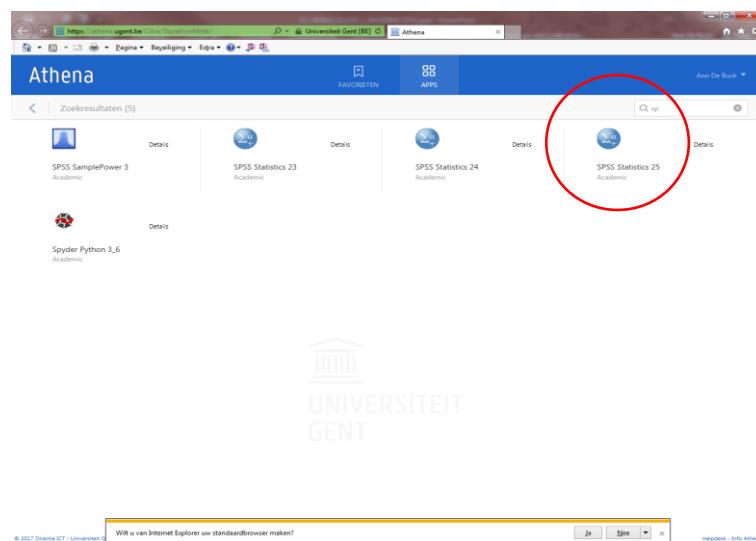
Van het programma SPSS bestaan verschillende versies bedoeld voor verschillende types van computers en gericht naar verschillende gebruikersgroepen. Wij maken gebruik van SPSS for Windows. De meest recente versie SPSS Statistics 27 is voor UGent studenten beschikbaar via Athena.

© 2002 Randy Glasbergen
www.glasbergen.com



3. SPSS OPENEN- OPSLAAN VAN GEGEVENS

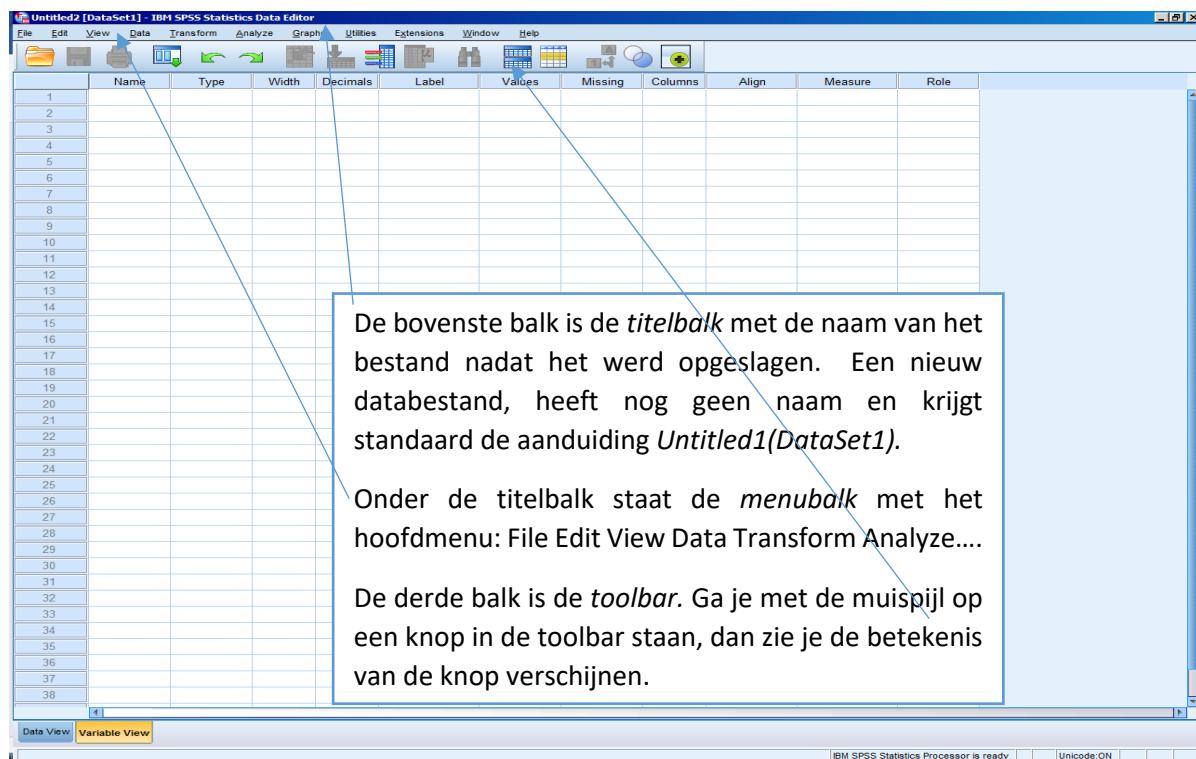
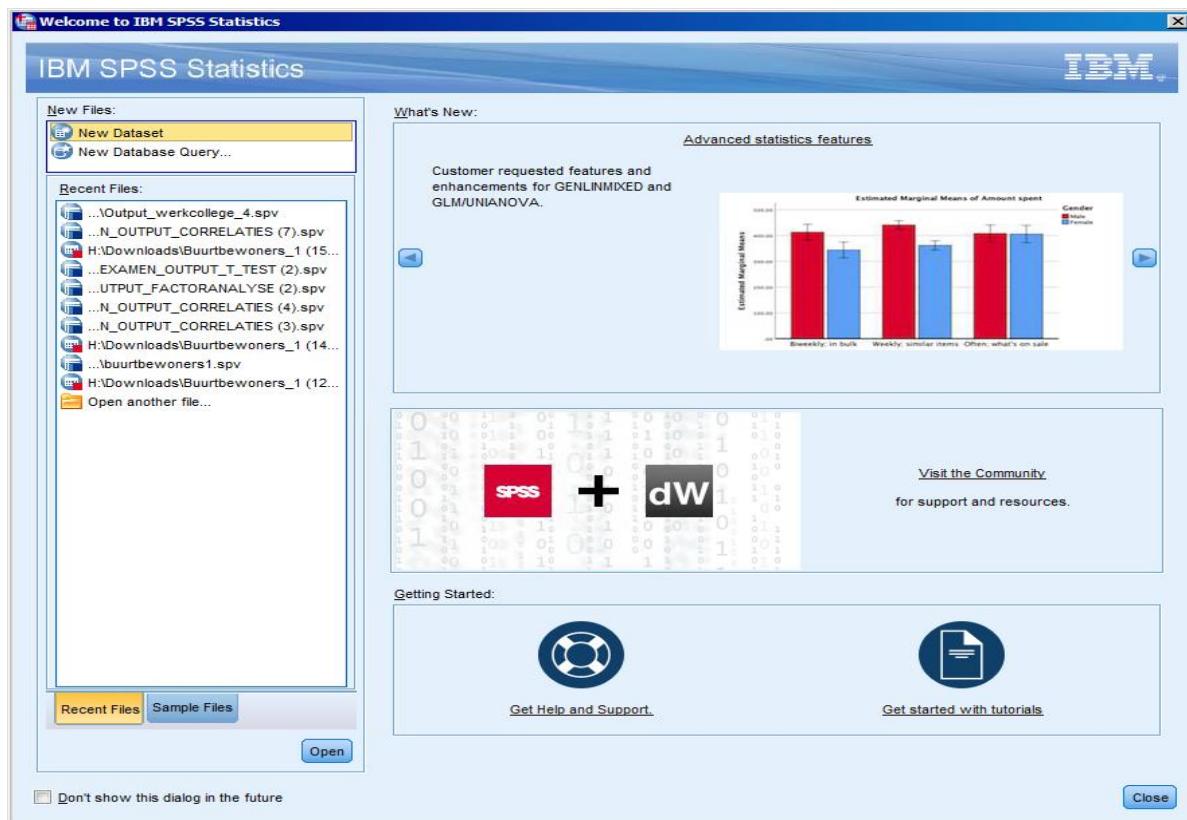
SPSS openen via Athena icoontje SPSS Statistics 27 en het programma start op.



Pop-up verschijnt met advies om data op te slaan als een .zsav bestand. Ok aanvinken.



Standaard toont SPSS de laatst gebruikte bestanden. Een ander bestand opvragen kan via *Open another file*. Een programma openen zonder databestand actief te maken, kan door *New dataset* aan te vinken.



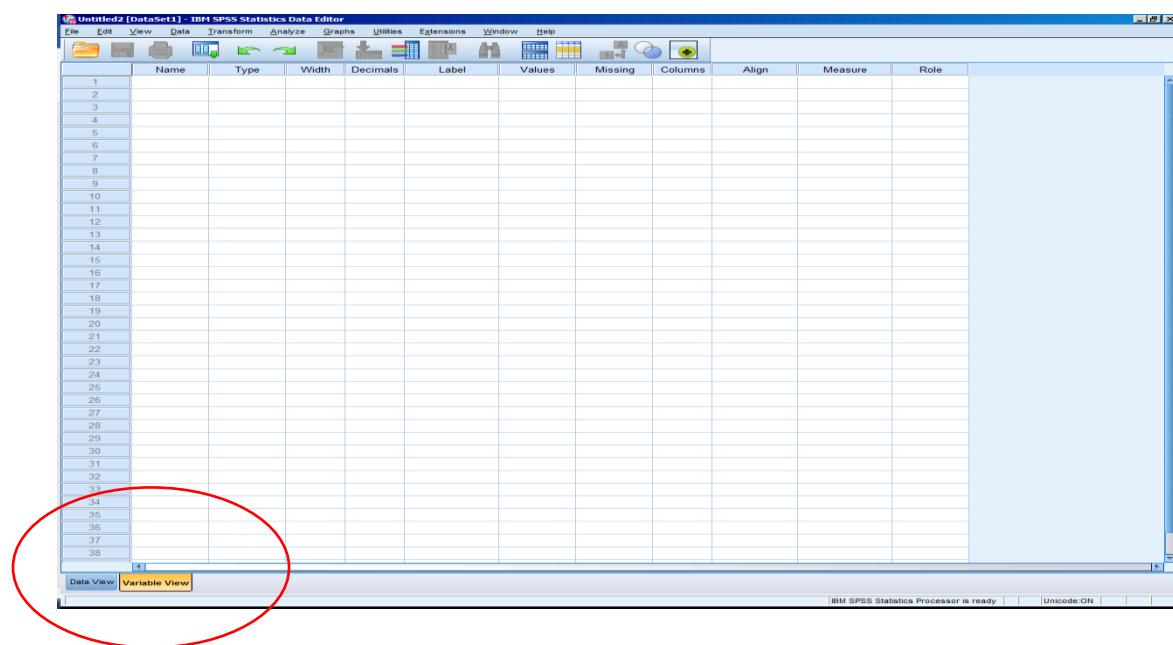
SPSS heeft verschillende *vensters* die gekoppeld zijn aan een apart type bestand dat je afzonderlijk moet openen en afzonderlijk moet opslaan.

Vensters openen: *File/Open/Data* *File/Open/Syntax* *File/Open/Output*

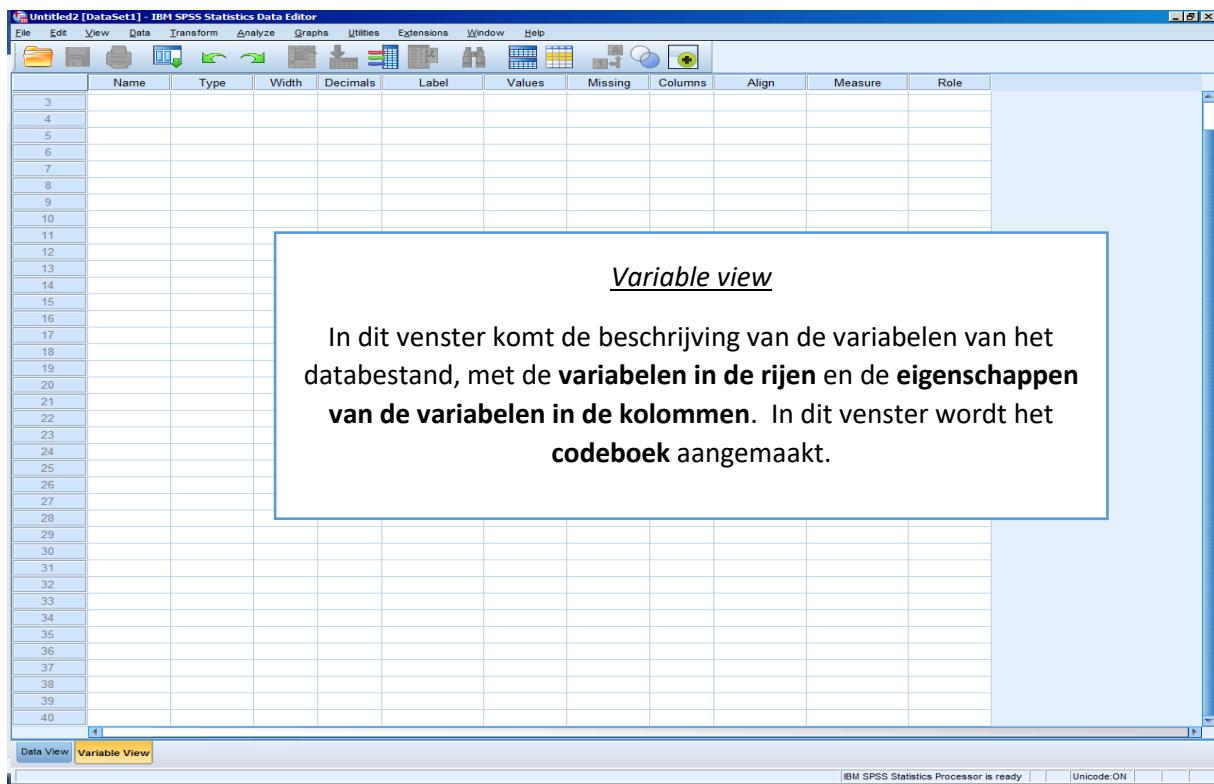
VENSTER	FUNCTIE INHOUD	BESTAND	EXTENSIE
DATA-editor	Hierin staan de data en variabelen	DATA	.zsav
SYNTAX-venster	Hierin sla je de syntax op van SPSS-opdrachten zodat je een eerdere opdracht nogmaals kunt laten uitvoeren. De syntax is de formule waarmee een statistische toets werd uitgevoerd	SYNTAX	.sps
OUTPUT-venster	Hierin komen de resultaten van de analyses en kun je alleen openen nadat je de analyses hebt gedaan	OUTPUT	.spo

Data-editor

De inhoud van de data-editor is de datamatrix of dataset. De dataset in SPSS bestaat uit **waarden (values)** van **statistische eenheden** of **onderzoekseenheden (cases)** (vb. individuen, maar ook aggregaten zoals scholen, buurten, landen, gevangenissen,...) op **variabelen (variables)**. Data-editor is onderverdeeld in twee deelvenster: *dataview en variable view*.

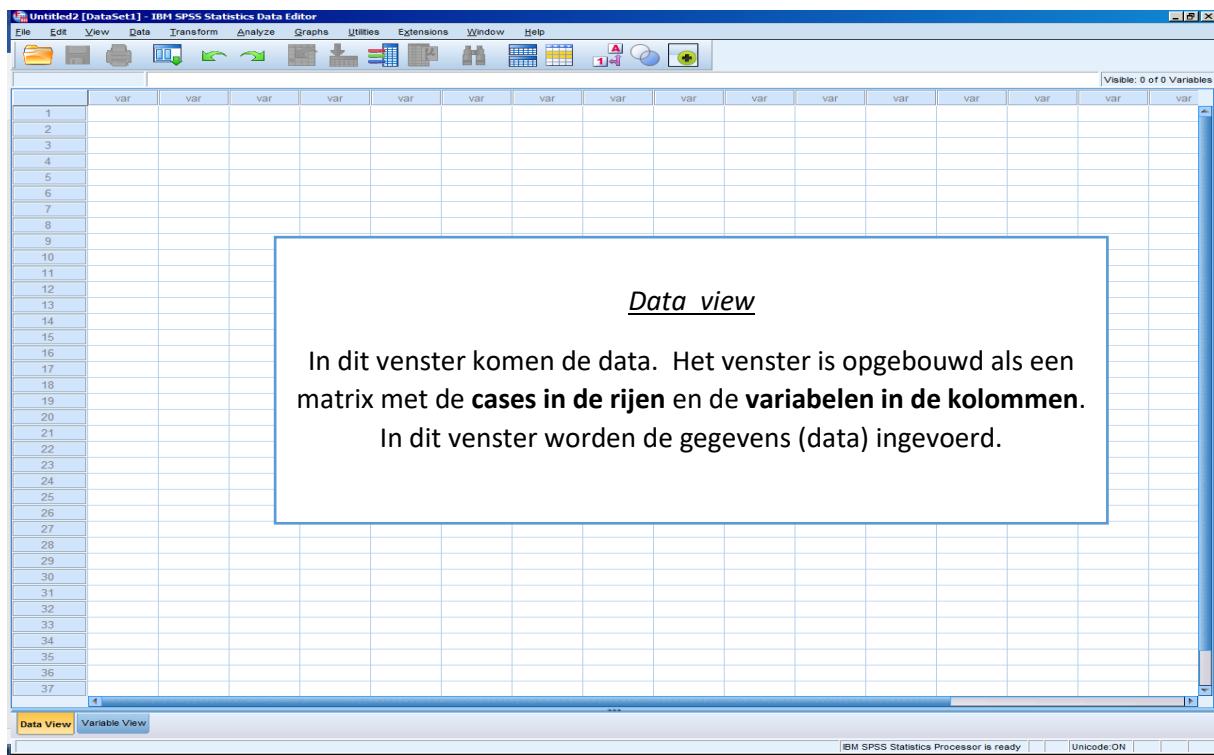


Verzamelde gegevens worden opgenomen in een gegevensbestand dat SPSS kan analyseren. Het SPSS-gegevensbestand is een matrix waarbij de waarnemingen in de rijen staan en de variabelen in de kolommen. Het maken van een gegevensbestand bestaat uit twee stappen: ten eerste het omzetten van de metingen in variabelen en ten tweede het invoeren van de gegevens in het gegevensbestand.



Variable view

In dit venster komt de beschrijving van de variabelen van het databestand, met de **variabelen in de rijen** en de **eigenschappen van de variabelen in de kolommen**. In dit venster wordt het **codeboek** aangemaakt.



Data view

In dit venster komen de data. Het venster is opgebouwd als een matrix met de **cases in de rijen** en de **variabelen in de kolommen**. In dit venster worden de gegevens (data) ingevoerd.

Toelichting bij Variable View

SPSS-VENSTER 1: VARIABLE VIEW

Toelichting bij kolommen in variable view	
Name	Naam van de variabele. Kies een kernwoord dat goed bij de variabele past of een vraagnummer. Bijv. <i>IDnummer</i> of <i>casenummer</i> als volgnummer voor de respondent of <i>V1_bandouders</i> (vraag 1 band met de ouders).
Type	Type variabele: - <i>numeriek</i> : variabele bestaat uit cijfers - <i>alfanumeriek (Eng: strings)</i> : variabele bestaat uit letters of andere tekens. Standaard staat het veld op numeriek.
Width	Het aantal karakters dat een ingevoerde waarde mag hebben. Bijv. 91,850 bestaat uit 6 karakters. Standaard staat het veld op Width 8: voor numerieke variabelen is dat meestal voldoende.
decimals	Het aantal decimalen van de ingevoerde waarde. Zijn er geen decimalen nodig, zet deze waarde dan op 0. Dit is overzichtelijker. Bijv. variabele 'gender' Standaard staat het veld op 2 decimalen.
Label	Uitgebreidere omschrijving voor de variabele. Hier kan je bijvoorbeeld de vraag uit de survey noteren. Handig om variabelen tijdens analyses beter te herkennen, vooral als meerdere onderzoekers op 1 dataset werken.

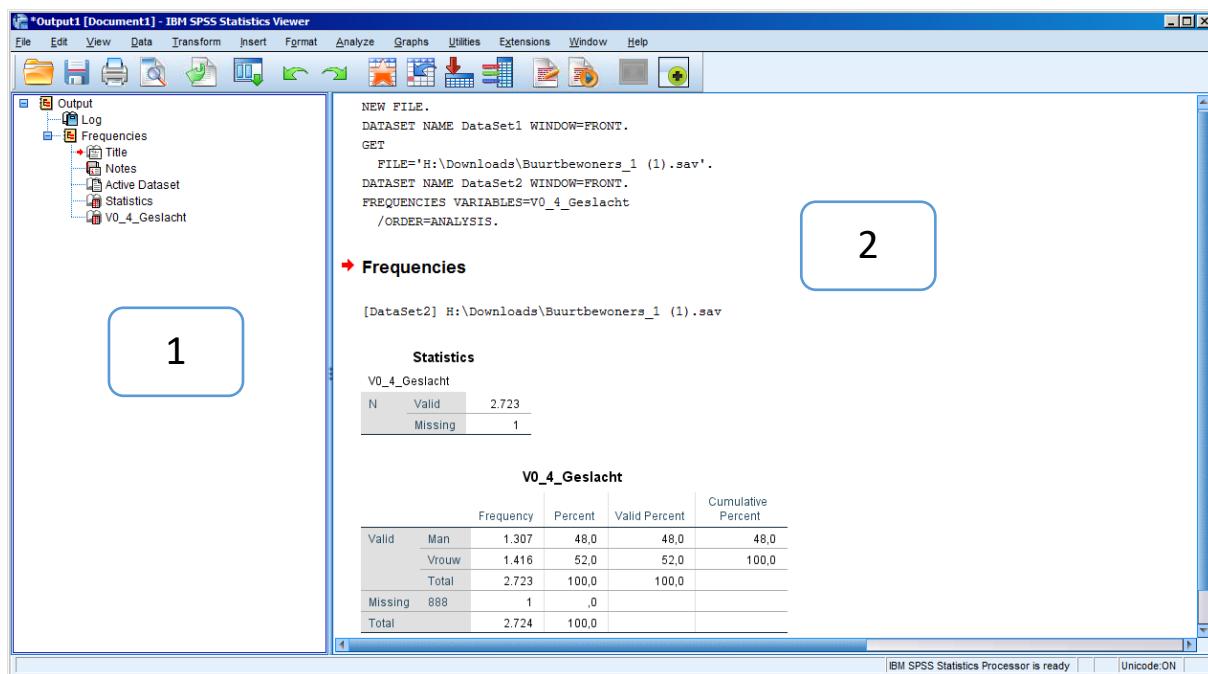
SPSS-VENSTER 1: VARIABLE VIEW

Toelichting bij kolommen in variable view	
Values	Codes die aan waarden van een variabele worden toegekend. Bijv. bij de variabele 'gender' staat code '1' voor jongens en code '2' voor meisjes. Standaard staat Values op None.
Missing	waarden die in het bestand ontbreken bijv. een respondent heeft een vraag niet ingevuld. Ofwel vul je hier niets in (no missing values) ofwel geef je aan de ontbrekende waarde een onmogelijke waarde zoals -999. Dit is ' <i>discrete missing value</i> '. Kies een waarde waarvan je onmiddellijk ziet dat het om een ontbrekende waarde gaat.
Columns	Kolombreedte van de variabele op het tabblad Data View Heeft geen invloed op de statistische analyse.
Align	Uitlijning van het tabblad Variable view Heeft geen invloed op de statistische analyse
Measure	De meetschaal van de variabele: nominaal, ordinaal of scale.

Output venster

In dit venster verschijnen de uitkomsten van de analyses in

1. Het **navigatievenster** : overzicht van de output. Klik op de uitkomst die je wilt zien.
2. Het **uitvoervenster**: hier lees je de uitkomsten van de analyses (grafieken en tabellen).



Zolang dit venster niet wordt afgesloten, worden de resultaten van opdrachten onder elkaar geplaatst in het openstaande outputvenster.

In het rechtergedeelte vinden we de output zelf terug. Resultaten van opeenvolgende analyses worden onder elkaar in dit venster weergegeven. In het linkergedeelte staat een navigatiestructuur. Die laat ons toe om snel van het ene naar een ander stuk van de output te springen. Stukken output kunnen ook tijdelijk verborgen worden door op het minteken te klikken bij de betreffende vermelding in de navigatiestructuur.

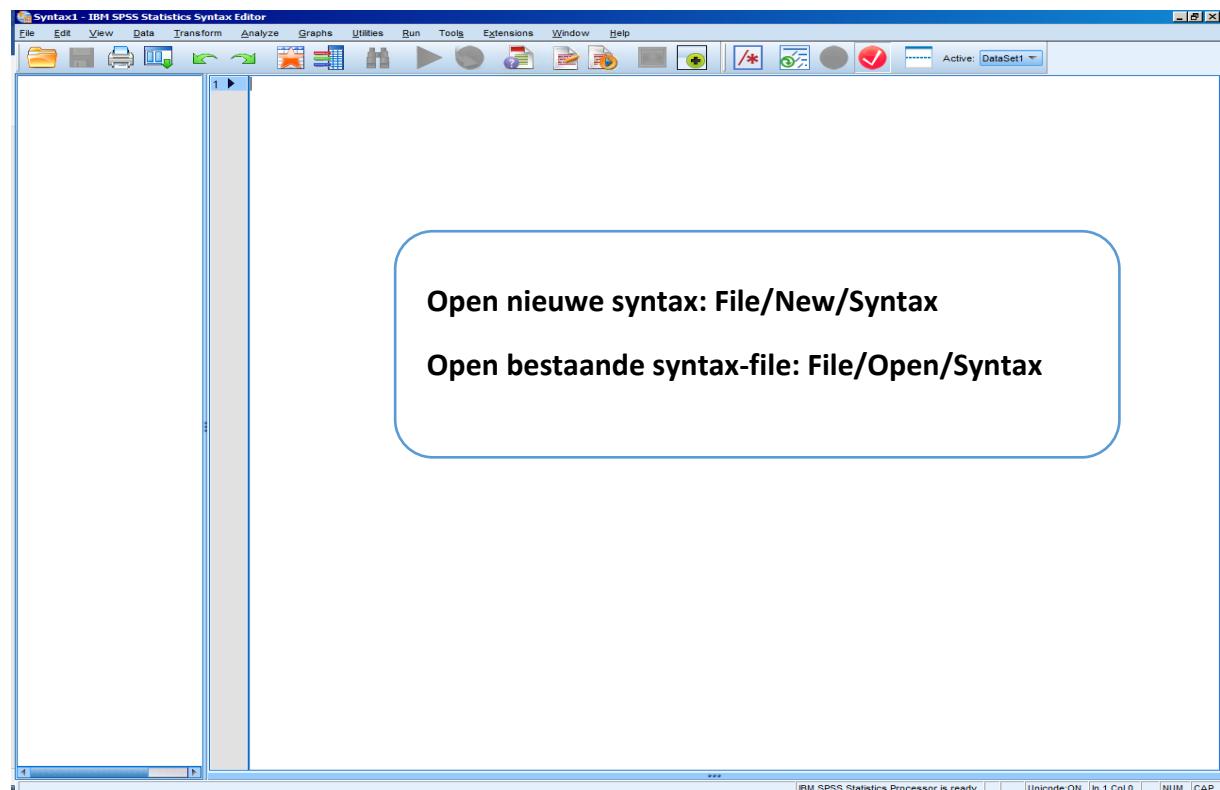
Boven in het output-venster vinden we de menubalk. We vinden er dezelfde elementen als in het data-editor-venster, met uitzondering van de items 'data' en 'transform'. Het bevat daarentegen wel twee nieuwe items:

- **Insert:** Hiermee kunnen bepaalde elementen (zoals een titel) worden toegevoegd aan de output.
- **Format:** Hiermee kan bij het printen de positie worden gewijzigd van de elementen in de output.

De lay-out van elementen in de output kan aangepast worden door op het betreffende element te dubbelklikken.

Syntax venster

Om instructies te geven door commando's in te tikken (in plaats van keuzes te maken uit de menu's) staat een syntax-venster ter beschikking.



Waarom syntax gebruiken ?

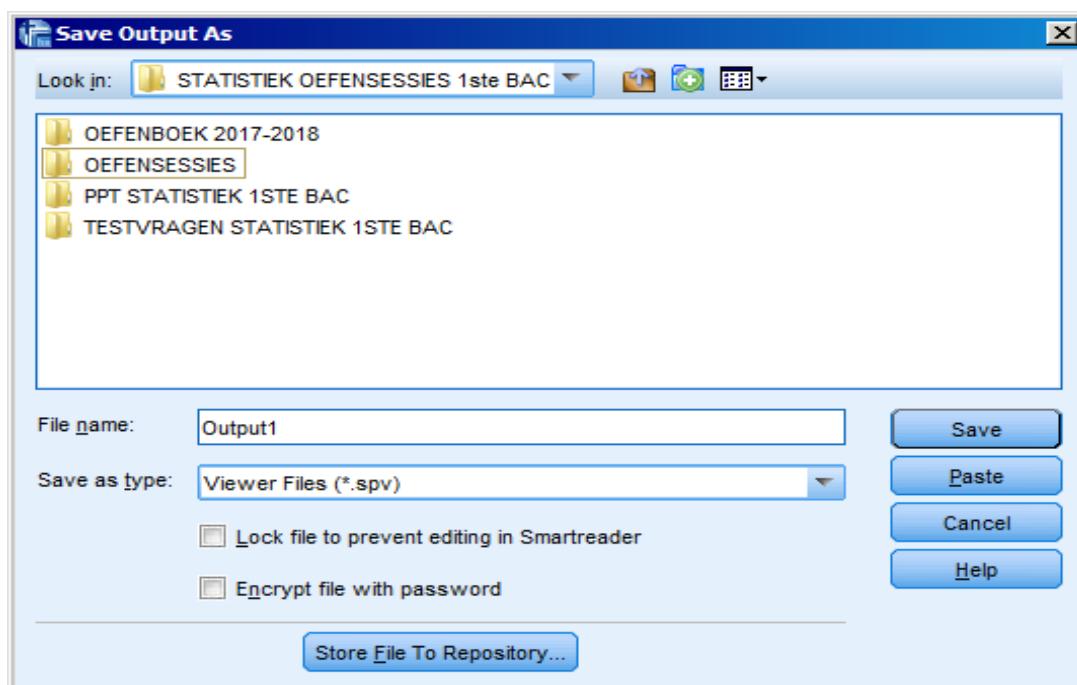
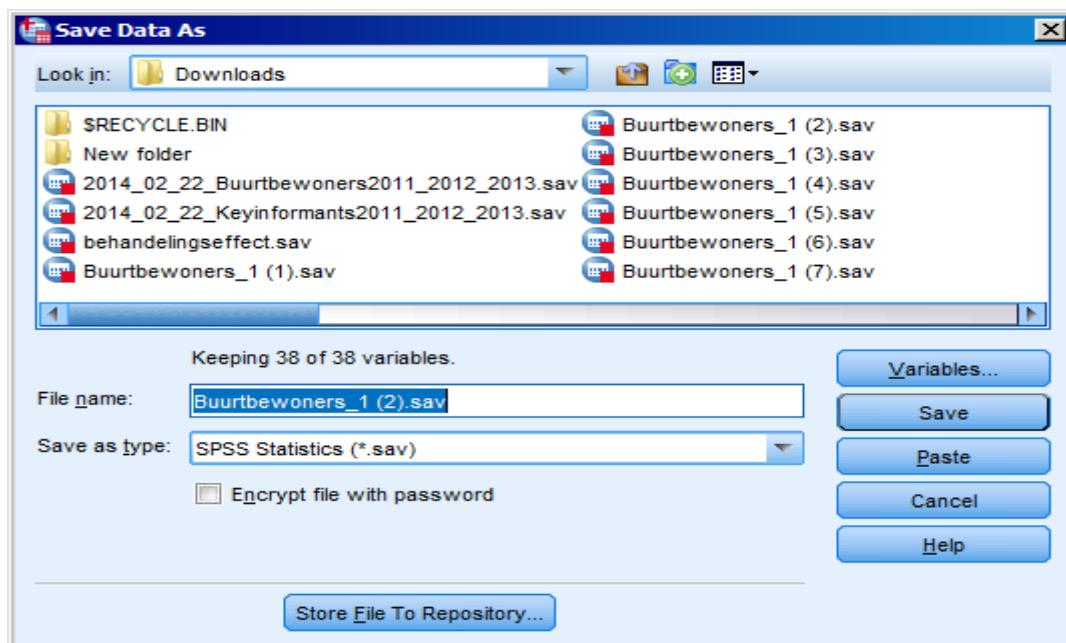
- **Herhaling van analyses**: wil je een analyse herhalen met andere variabelen, dan is het eenvoudiger om de syntax aan te passen. Anders moet je alle dialoogvensters en submenu's terug instellen
- **Bewaren**: vaak worden veel analyses na elkaar uitgevoerd en weet je niet meer welke instructies zijn gegeven. In een afzonderlijk opgeslagen syntaxbestand staan alle commando's netjes onder elkaar zodat raadplegen en reconstructie steeds mogelijk is.

In het onderste gedeelte kunnen de commando's ingetikt worden. Die commando's kunnen uitgevoerd worden door in het menu *Run* te kiezen voor: All (voert alle commando's uit), Selection (voert de geselecteerde commando's uit), Current (voert commando uit waarop de cursor staat), of To end (voert commando uit waarop de cursor staat en alle volgende commando's).

Opslaan van gegevens

Wil je je werk bewaren dan moet je de ingevoerde gegevens plus de output opslaan.

- **Opslaan gegevens (data) in data-editor**: ga naar *File/save as* en kies het bestand waar je wilt opslaan; **Opgelot : sla op als .zsav**. Klik op save. De tabbladen Data View en Variable View zijn beide opgeslagen in 1 bestand.
- **Opslaan van de output**: herhaal bovenstaande procedure in het outputvenster. Outputfile wordt altijd opgeslagen met de extensie .spv. Klik op save.



Vooraleer een SPSS-sessie te beëindigen moet steeds de bedenking worden gemaakt of alles wat van belang kan zijn voor later gebruik, op schijf werd bewaard. Als SPSS wordt beëindigd zonder dat dit gebeurde, kan belangrijke informatie verloren gaan waardoor reeds verrichte analyses moeten worden overgedaan. Dit kan zowel gaan om databestanden waarin wijzigingen werden aangebracht, als om de output die op het scherm verscheen. Als de output van belang is, moet die worden bewaard. **SPSS vraagt voor ieder geopend venster met nog niet bewaarde wijzigingen of de inhoud ervan op schijf moet worden bewaard.**

4. INVOEREN VAN GEGEVENS

Aan de hand van onderstaand voorbeeld kun je zelf een mini-databestand opbouwen.

ZELF AAN DE SLAG

FICTIEF VOORBEELD: SURVEY		
Respondentnummer_____		
Gender:	M V	Leeftijd: _____ (in jaren) Geboortejaar: _____
Faculteit:		
1. Recht en Criminologie		
2. Letteren en Wijsbegeerte		
Op een schaal van 1 tot 4, hoe zeker ben je dat je voor het opleidingsonderdeel Toegepaste data-analyse zal slagen ?		
1 absoluut niet zeker 2 niet zeker 3 vrij zeker 4 absoluut zeker		
Hoeveel jaar ervaring met SPSS heb je al ? (omcirkel slechts 1 antwoord)		
<input checked="" type="radio"/> 0 geen ervaring <input type="radio"/> 1 1 jaar ervaring <input type="radio"/> 2 2 jaar of meer ervaring		

18

STAP 1: creatie van het codeboek/aanmaken van variabelen

Start SPSS op en ga naar data-editor, tabblad Variable View.

Alle vragen uit het fictieve survey-voorbeeld peilen naar een kenmerk van de respondent en worden beschouwd als een variabele. In bovenstaand voorbeeld is er sprake van 7 variabelen.

Bedenk hoe je de variabelen in SPSS wilt benoemen, blijf zo dicht mogelijk bij de oorspronkelijke vraag.

Op de eerste rij in het tablad Variable View komt het respondentnummer of IDnummer. Typ dit in onder Name en druk op enter. In de overige kolommen zal SPSS automatisch een standaardwaarde invullen. Ga verder naar rij 2 en voer gender of geslacht in.

Rij 3: leeftijd, rij 4: geboortejaar, rij 5: faculteit, rij 6: perceptie_slaagkans en rij 7: ervaring_SPSS.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	IDnummer	Numeric	8	0	nummer respondent	None	.999	8	Right	Nominal
2	gender	Numeric	8	0		(1. jongens)...	-.999	8	Right	Nominal
3	leeftijd	Numeric	8	0	leeftijd in jaren	None	.999	8	Right	Scale
4	geboortejaar	Numeric	8	0		None	-.999	8	Right	Scale
5	faculteit	Numeric	8	0		{1, Recht en Criminologie}...	-.999	8	Right	Nominal
6	perceptie_slaagka...	Numeric	8	0	hoe zeker ben je dat je zal slagen	{1, absolut niet zeker}...	-.999	8	Right	Scale
7	ervaring_SPSS	Numeric	8	0	hoeveel jaar ervaring met SPSS	{0, geen ervaring}...	-.999	8	Right	Scale
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										
30										
31										
32										
33										
34										
35										
36										
37										
38										

Verder invullen van de kolommen:

- Type** staat standaard door SPSS ingevuld als *Numeric*
- Width** staat standaard op '8'. Je kan dit veranderen door in de cel te klikken
- Decimals** staat standaard op '2'. Tip: zet dit op 0 om alles overzichtelijk te houden

Label

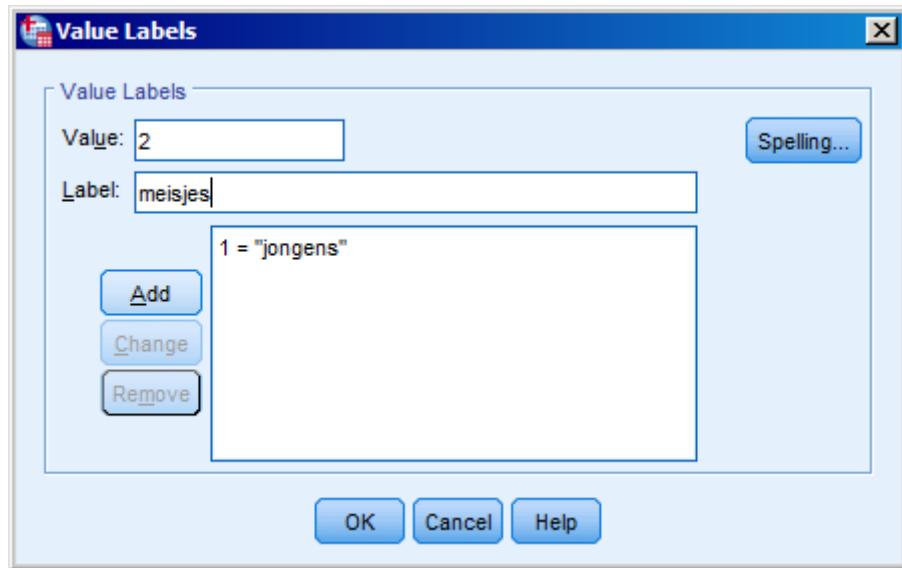
hoeft niet ingevuld om analyses uit te voeren maar geeft duidelijkheid. Hier kan je bijvoorbeeld de vraag uit de vragenlijst invoeren zodat je achteraf variabelen gemakkelijker kunt herkennen/opzoeken.

Values

Hoewel het in SPSS ook mogelijk is om te werken met niet-numerieke waarden van variabelen (in de voorbeeldtabel: de naam en het geslacht van de respondent) is het in de praktijk vaak makkelijker om niet-numerieke waarden door numerieke codes te vervangen (bvb. voor GESLACHT: '1' voor vrouwen en '0' voor mannen). Dit vereenvoudigt het intikwerk bij het ingeven van een datamatrix en vermindert de kans op fouten daarbij. **Let op:** Dit houdt echter in dat de numerieke score op zo een variabele eigenlijk niet meer is dan een code, die niet mag geïnterpreteerd worden als de echte waarde. De code dient immers alleen om de verschillende categorieën van elkaar te onderscheiden. Omdat de gebruikte codes in zo een geval niet erg betekenisvol zijn, bestaat de mogelijkheid om aan elke score een 'label' (beschrijving) toe te kennen die de echte waarde weergeeft en die in de output van de analyses wordt vermeld.

Pas vanaf het intervalniveau zijn de toegekende scores ook de echte waarden. Dat alle scores numeriek zijn, houdt ook in dat SPSS bereid is om alle analysetechnieken met om het even welke veranderlijken uit te voeren. Zo zouden we aan SPSS zonder problemen kunnen vragen om het gemiddelde geslacht te berekenen, hoewel dit onzin is. Het is dan aan de gebruiker om het meetniveau van de variabelen waarmee hij werkt te kennen en aan de hand daarvan de geschikte procedures te selecteren.

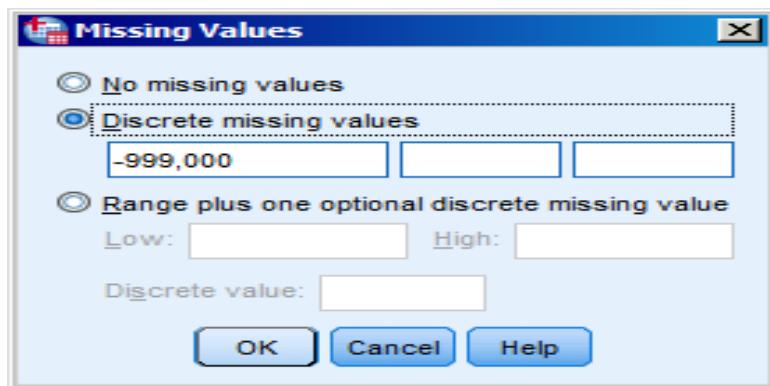
Klik in de kolom Values op de cel van de variabele *gender* en vervolgens op het blauwe vierkantje. Hiermee kom je het venster **Value Labels**. Type in het vak achter Value: 1 en achter Label: jongens. Klik op Add en herhaal voor meisjes (value: 2 en label: meisjes).



Geef ook de variabelen *faculteit*, *perceptie_slaagkans* en *ervaring_SPSS* een code. Gebruik bij het coderen van de categorieën van de variabelen dezelfde numerieke waarden als in de vragenlijsten.



Missing



Respondenten beantwoorden niet steeds alle vragen. Heeft een respondent nagelaten een vraag te beantwoorden, dan is het vaak handig om een waarde te kiezen voor het missende antwoord. Kies, om verwarring te voorkomen, een waarde waarvan je meteen weet dat het om een ontbrekende waarde gaat.

Er zijn twee soorten missing values:

- **User missing values:** is een respondent vergeten om een vraag te beantwoorden, dan kan hier een fictief getal ingevuld worden dat buiten de range van de variabele ligt (vb. 999)
- **System missing values:** waarden zelf door SPSS gecreëerd wanneer een case of een cel niet is ingevuld en weergegeven door een puntje

Columns

staat standaard op 8 en is over het algemeen voldoende

Align

staat standaard op *left* bij stringvariabelen en op *right* bij numerieke

Measure

geef aan om welk type variabelen het gaat: nominaal, ordinaal of scale

STAP 2: het invoeren van data in Data View

Als het codeboek in Variabele View is aangemaakt, kunnen de data werkelijk worden ingevoerd in het tabblad Data View. (klik linksonder in het tabblad Variabele View).

In het tabblad Data View staan de aangemaakte variabelen in de kolommen en de rijen komen de respondenten met daarachter de antwoorden op de respectievelijke variabelen.

Hieronder vind je fictieve data voor 4 respondenten.

ZELF AAN DE SLAG: INVOEREN DATA

Casenummer	Gender	Leeftijd	Geboortejaar	Faculteit	Slaagkans 1 tot 4	Ervaring SPSS
1	Jongen	18	1999	Recht en Crim	Absoluut zeker	Geen
2	Jongen	20	1997	Letteren & Wijsbegeerte	Absoluut niet zeker	Geen
3	Jongen	17	Geen antwoord	Recht en Crim	Vrij zeker	1jaar
4	meisje	18	1999	Recht en Crim	Absoluut zeker	2jaar

- Ga naar tabblad *data view* en vul datagegevens uit tabel hierboven in.

The screenshot shows the SPSS Data Editor window with the following data:

IDnummer	gender	leeftijd	geboortejaar	faculteit	perceptie_slaagkans	ervaring_SPSS
1	1	18	1.999	1	4	0
2	1	20	1.997	2	1	0
3	1	17	-999	1	3	1
4	2	18	1.999	1	4	2
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31						
32						
33						
34						
35						

Tip ! Maak bij het invoeren van datagegevens gebruik van 'value labels' in toolbar

5. HET BEWERKEN VAN WAARNEMINGEN

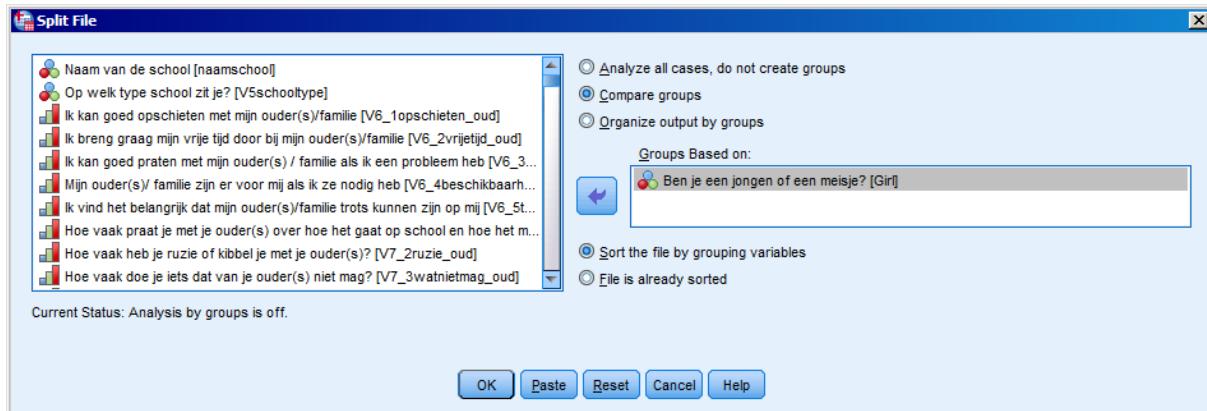
Met bewerkingsprocedures kunnen in SPSS:

- groepen opgesplitst worden voor identieke analyses per groep : **Split File**
- waarnemingen geselecteerd worden : **Select Cases**
- hercoderen van variabelen : **Recode into Same/Different Variables**

Split File

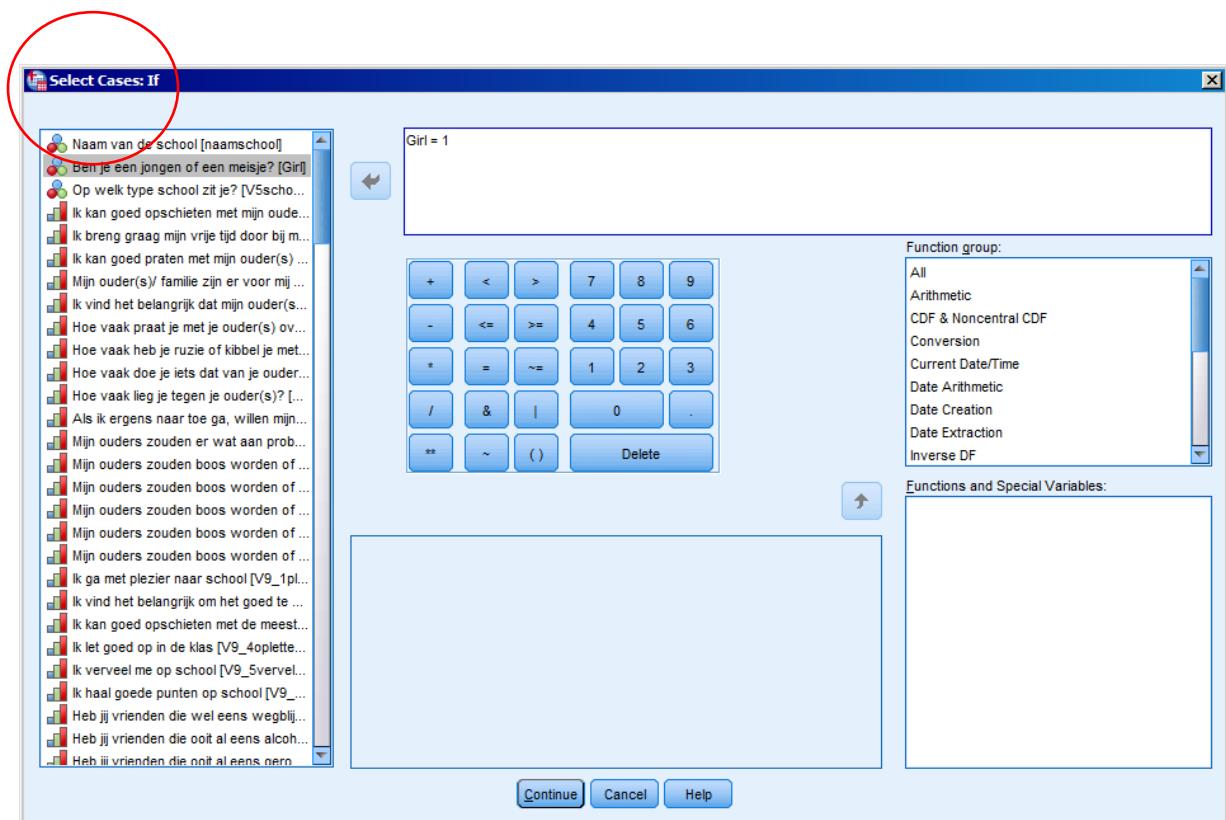
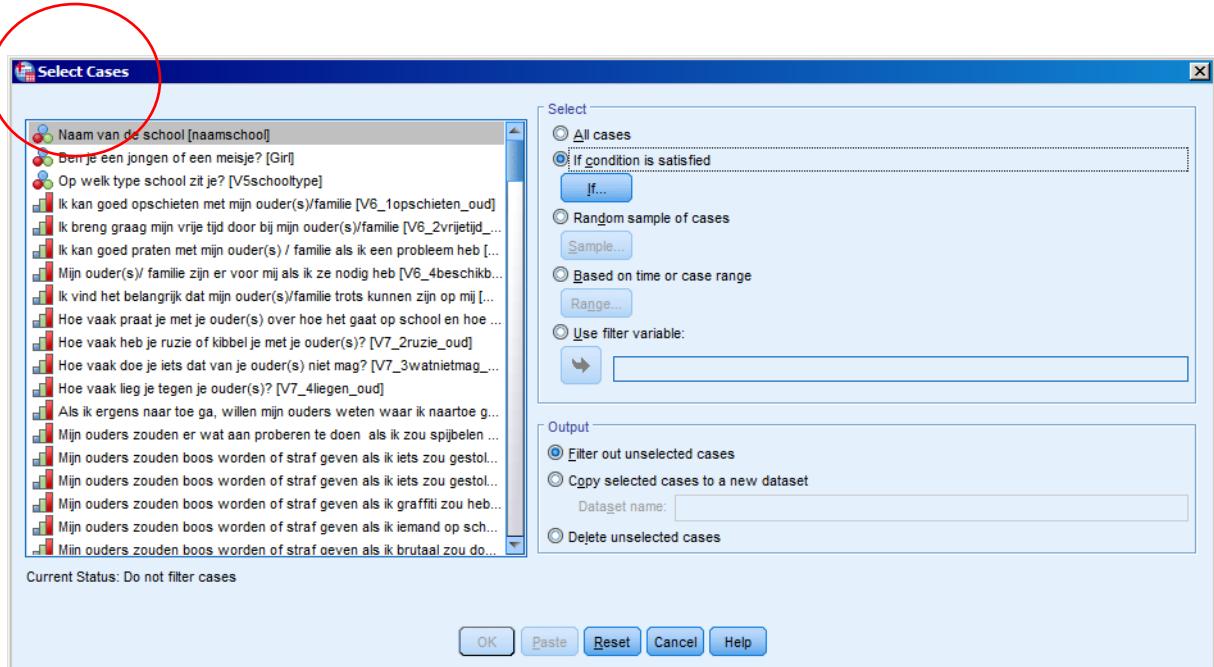
Met de opdracht Split File kan een groepsindeling opgegeven worden in SPSS waarna elke volgende analyse automatisch voor elke groep afzonderlijk wordt uitgevoerd.

Stel dat we de beschrijvende statistieken voor jongens en meisjes apart willen presenteren, dan geven we via de opdracht *Data / Split File* een groepsdeling op (*compare groups of via organize output by groups*). Alle analyses die we daarna uitvoeren, wordt voor elke groep afzonderlijk uitgevoerd.



Select Cases

De opdracht Select cases selecteert een bepaalde groep waarnemingen waarna de volgende analyses alleen voor deze groep waarnemingen worden uitgevoerd. Waarnemingen kunnen geselecteerd worden op basis van een voorwaarde, toeval of case-nummer. Bijvoorbeeld: via de opdracht *Data / Select Cases / If condition is satisfied / variabele 'geslacht' = 1* kunnen alle meisjes in de dataset worden geselecteerd (op voorwaarde dat meisjes als '1' werden gecodeerd). Met Select Cases kan ook een aselecte steekproef worden getrokken. Willen we enkel de 50 eerste waarnemingen betrekken in onze analyses, dan kan dit ook via Select Cases.



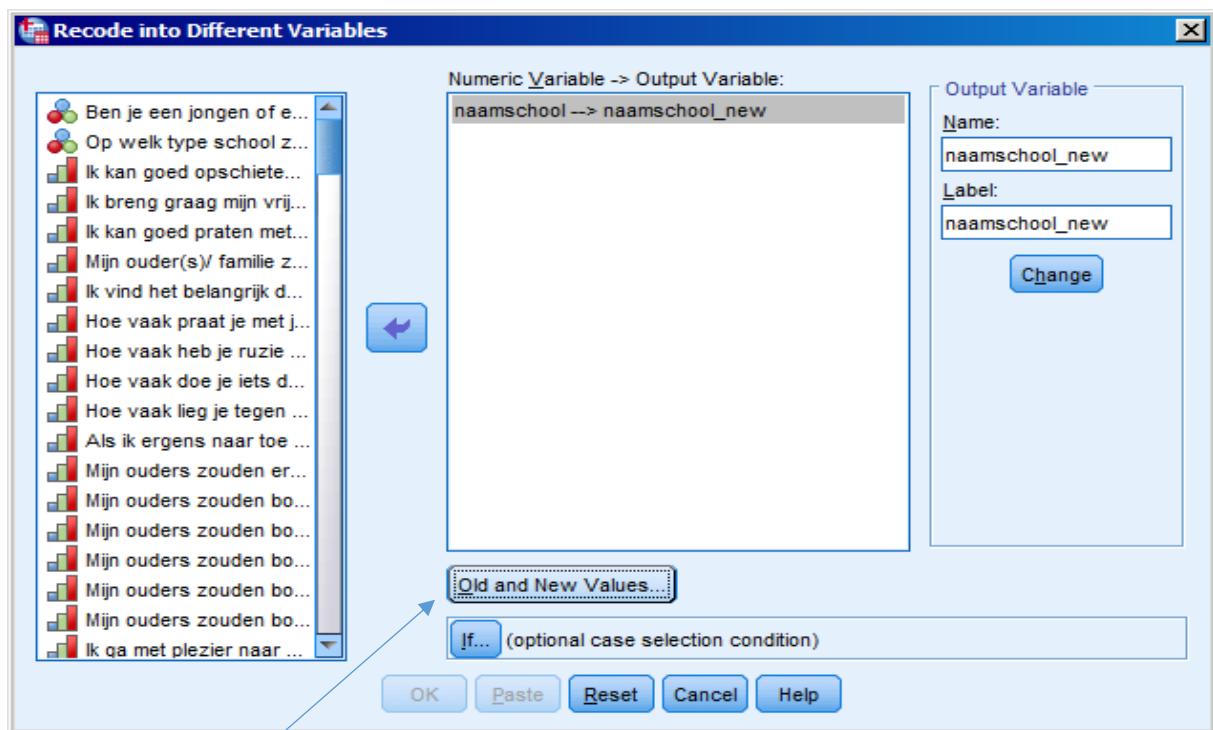
Opgelet ! zolang de opdracht in SPSS niet werd ‘geset’, zullen alle volgende analyses uitgevoerd worden voor groepen apart of geselecteerde waarnemingen.

Recode into Same/Different Variables

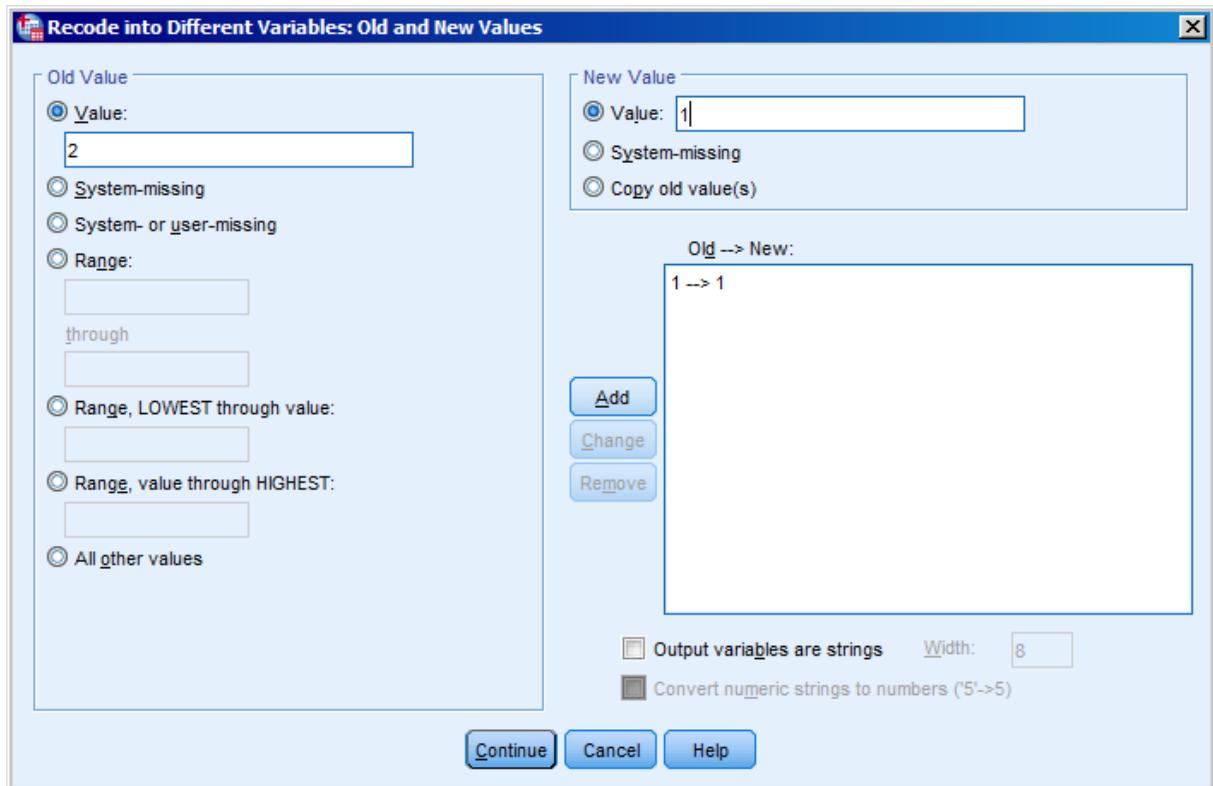
Via de opdracht *Recode into Same/Different Variables* (onder menuknop *Transform*) kunnen antwoordcategorieën van variabelen hercodeerd worden. Bijvoorbeeld: we willen van een variabele met 4 antwoordcategorieën (1 tot 4) een dichotome variabele maken waarbij we de waarden 1 en 2 gaan hercoderen naar waarde 1 en waarden 3 en 4 gaan hercoderen naar waarde 2.

Recode into Same Variable verandert een oorspronkelijke variabele naar een nieuwe variabele, met *Recode into Different Variable* behoud je de oorspronkelijke variabele en wordt een nieuwe variabele gecreëerd. We raden aan om bij voorkeur deze laatste optie te gebruiken. Op die manier vermijd je dat oorspronkelijke variabelen verloren gaan.

Selecteer de variabele die je wilt hercoderen, geef onder *output variable* een nieuwe naam en label, klik op *change*.



Onder *Old and New Values* geven we aan de oude (originele) waarden een nieuwe waarde.



Door op *Add* te klikken wordt de hercodering als uit te voeren actie toegevoegd. Klik op *Continue* en *OK*. De bestaande variabele wordt nu vervangen door een nieuwe met nieuwe codes. Vergeet niet om onder *Values* de categorieën van de nieuwe variabele aan te geven.

6. UITVOEREN VAN STATISTISCHE ANALYSES

In deze kennismaking met SPSS gaan we kort in op een aantal basis analysetechnieken. In hetgeen volgt komen volgende analyses aan bod:

1. Univariate beschrijvende statistieken
2. Samenhang tussen twee variabelen
3. Verklaren van een variabele op basis van één of meerdere onafhankelijke variabelen

6.1.Univariate beschrijvende statistieken

Het opvragen van frequentietabellen, parameters van centraliteit, spreiding en vorm gebeurt in SPSS via *Analyze Descriptive Statistics Frequencies*. Frequenties kunnen getoond worden in een tabel of een grafiek (*Charts*).

Hierna tonen we hoe je de univariate beschrijvende statistieken opvraagt voor nominale, ordinale en metrische variabelen. Eerst wordt een woordje uitleg gegeven bij het ingeven van instructies in dialoogvensters in SPSS.

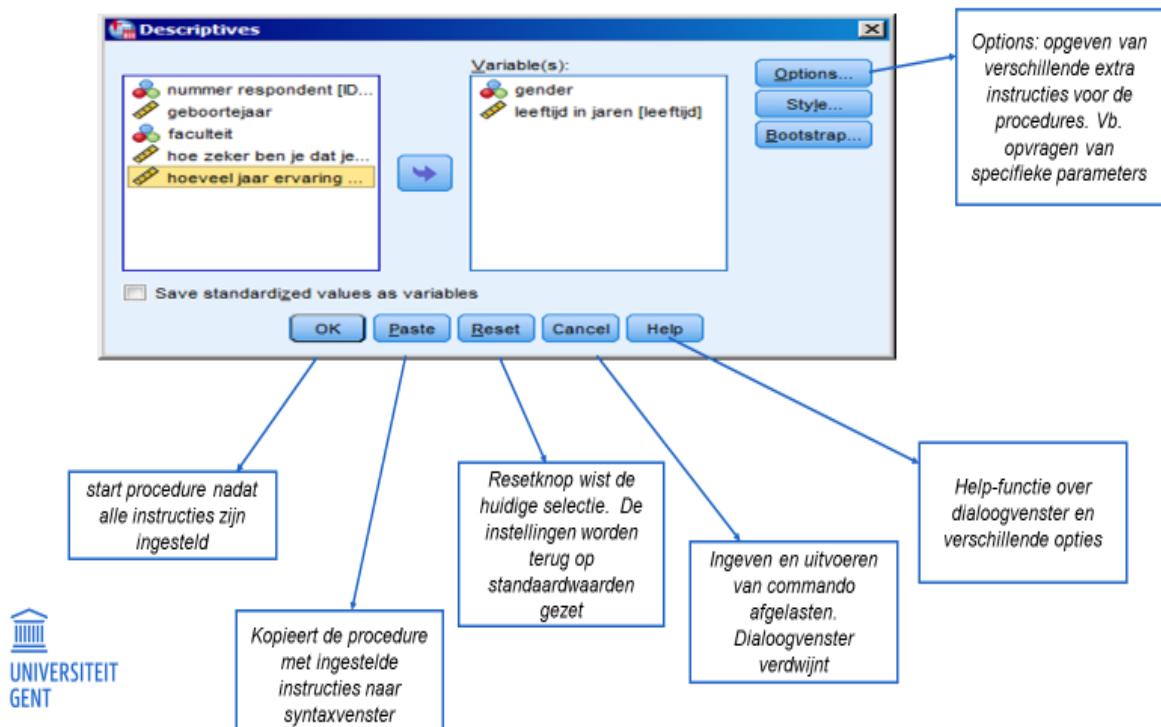
Ingeven van instructies in dialoogvensters

Twee werkwijzen kunnen gevuld worden om SPSS aan te zetten om instructies te geven: kiezen van opties in de menustructuur of intikken van commando's in een syntax-venster. Het uitvoeren van commando's via syntax betekent dat we een bepaalde commandotaal gebruiken om SPSS berekeningen te laten doen. We behandelen in deze inleidende syllabus enkel het werken met de menustructuur.

In alle SPSS-dialoogvensters vinden we volgende commandoknoppen:

DIAЛОOGVENSTERS IN SPSS: VOORBEELD

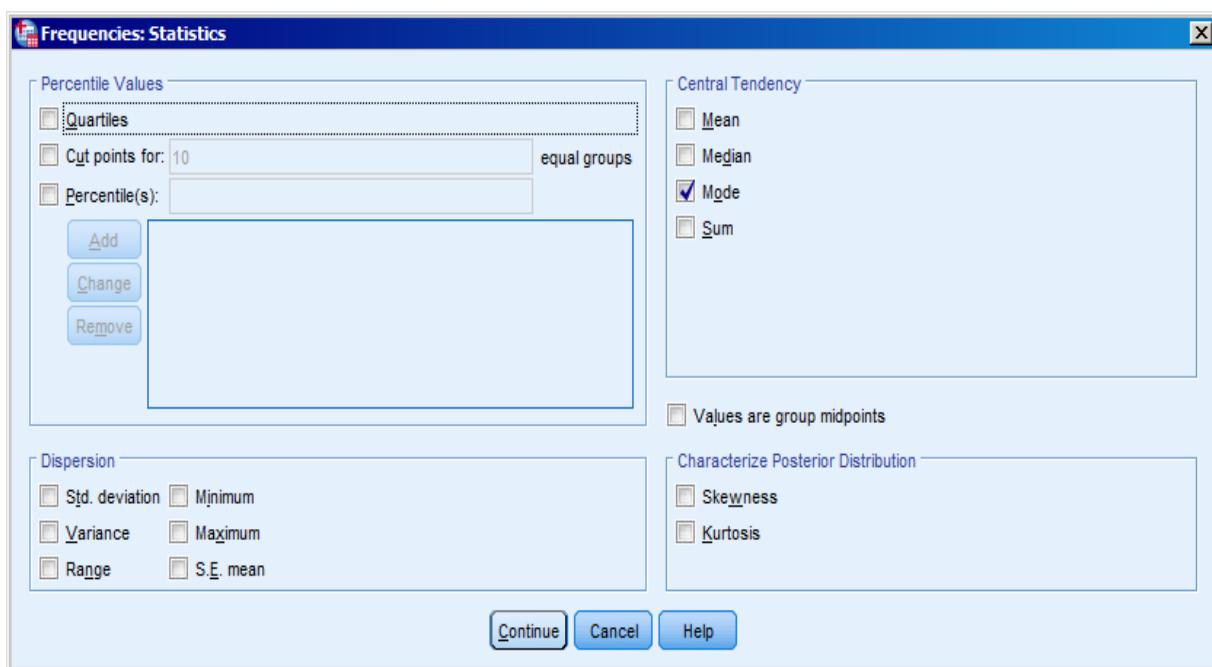
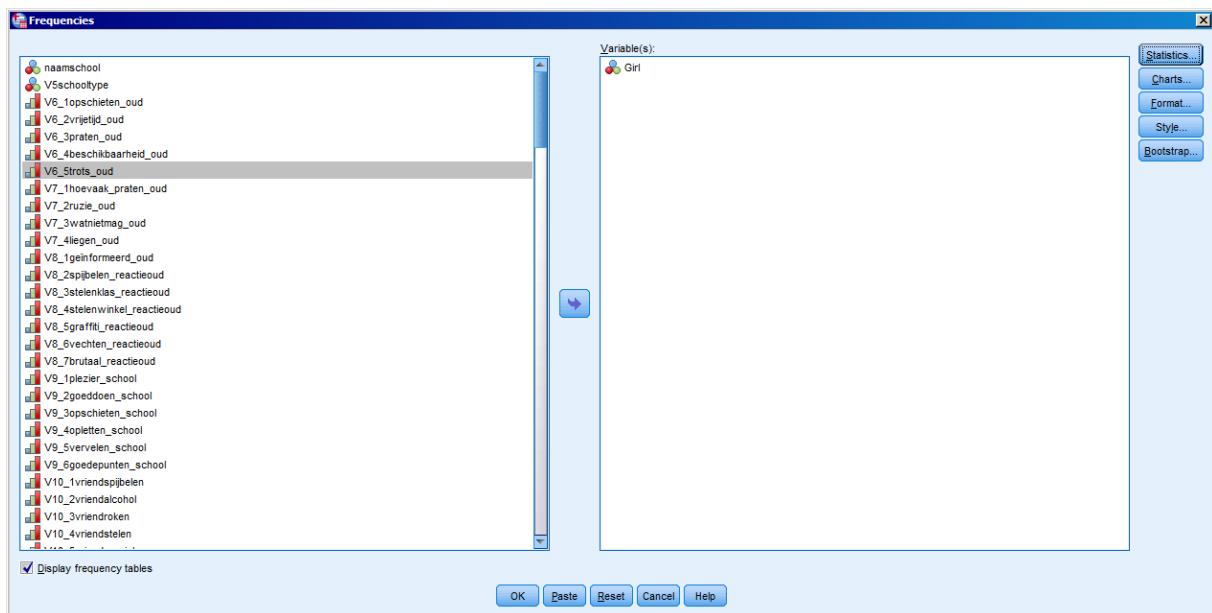
- Selecteer gender en leeftijd



Let op! De selecties die gemaakt werden in een dialoogvenster blijven gedurende een hele SPSS-werksessie onveranderd, tenzij die door de gebruiker zelf worden gewijzigd. Dit noemt men '**persistentie van de instellingen**' en heeft tot gevolg dat een gebruiker steeds goed moet nakijken welke opties aan en uit staan, en of dit inderdaad overeenkomt met de wensen van de analyse. De 'Reset'-toets zorgt ervoor dat de instellingen van een vorig gebruik van een dialoogvenster worden gewist.

Parameters van centraliteit en spreiding voor nominale variabelen

Ga naar Analyze Descriptive Statistics Frequencies. Kies onder Statistics Mode Continue OK



Ben je een jongen of een meisje?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	jongen	386	47,4	47,9	47,9
	meisje	420	51,5	52,1	100,0
	Total	806	98,9	100,0	
Missing	System	9	1,1		
	Total	815	100,0		

Valid categorieën van de variabele met de geldige waarden

Missing item-nonrespons of ontbrekende waarden

Frequency absolute frequenties

Percent absolute frequenties gepercenteerd

Valid percent geldige percentages

Cumulative percent cumulatieve, geldige percentages

In een zelfrapportage survey werd aan 815 leerlingen uit de basisschool gevraagd aan te duiden of zij een jongen of een meisje waren. Het meetniveau van deze variabele is **nominaal**. De antwoordcategorieën zijn ‘jongen’ en ‘meisje’. In totaal hebben 806 leerlingen de vraag beantwoord, 9 leerlingen deden dat niet. Er zijn 9 missings (1.1%). De analyse is dus gebaseerd op 98.9% (806 respondenten) van de respondenten.

Statistics

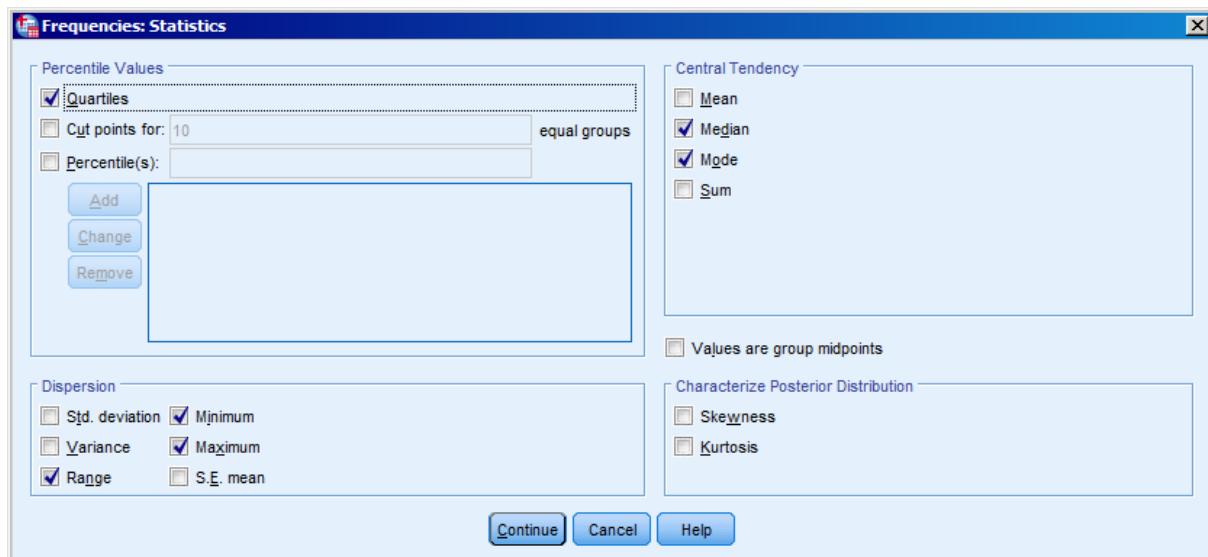
Ben je een jongen of een meisje?

N	Valid	806
	Missing	9
Mode		1

Bekijken we de parameter van centraliteit: de modus. De modale waarde bedraagt 1. Code 1 staat voor meisjes (jongens werden gecodeerd met code ‘0’). Dit betekent dat de meeste respondenten meisjes zijn. En wat met de spreiding? We kunnen kijken naar het percentage dat niet tot de modus behoort, of we kunnen de variatieratio, de index van diversiteit of de spreidingsmaat ‘d’ zelf berekenen.

Parameters van centraliteit en spreiding voor ordinale variabelen

Ga naar Analyze Descriptive Statistics Frequencies. Kies onder Statistics Quartiles Mode Median Range Minimum Maximum Continue OK



Hoe vaak lieg je tegen je ouder(s)?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	(bijna) nooit	515	63,2	63,3	63,3
	paar keer per maand	219	26,9	26,9	90,3
	paar keer per week	68	8,3	8,4	98,6
	(bijna) elke dag	11	1,3	1,4	100,0
	Total	813	99,8	100,0	
Missing	System	2	,2		
	Total	815	100,0		



We vroegen aan 815 leerlingen uit de basisschool ‘*hoe vaak lieg je tegen je ouders?*’ De antwoordcategorieën zijn ‘(bijna) nooit’ (code 0), ‘een paar keer per maand’ (code 1), ‘een paar keer per week’ (code 2), ‘(bijna) elke dag’ (code 3). We kunnen in deze situatie de respondenten ordenen op een niet-nauwkeurig gedefinieerde schaal. De afstanden tussen de categorieën zijn immer niet geijkt.

We interpreteren deze tabel als volgt. Een overgrote meerderheid van de leerlingen (63.3 % of 515 van de bevraagde leerlingen) zegt (bijna) nooit te liegen tegen de ouders. Niettemin zegt 26.9% (of 219 respondenten) dat toch een paar keer per maand te doen. Samen maken deze 2 categorieën 90.3% van de geldige respons uit. Dit betekent dat er slechts heel weinig leerlingen aangeven een paar keer per week of (bijna) elke dag te liegen tegen de ouders. Wat zegt deze tabel nu inhoudelijk voor de criminoloog-onderzoeker? Uiteindelijk gaat het in onderzoek steeds om de inhoudelijke informatie die voortkomt uit de tabel. Als criminoloog zien we verschillen en grote tendensen onder het fenomeen ‘liegen tegen de ouders’. De grote tendens is dat de meeste jongeren (bijna) nooit liegen tegen de ouders (tenminste dat rapporteren de leerlingen in de steekproef), niettemin zijn er **verschillen (variatie of spreiding)** waar te nemen. Het zijn precies die verschillen die ons later zullen interesseren. Criminologen willen hiervoor immers een verklaring bieden en deze toetsen. Op dit ogenblik krijgen we enkel zicht op de aard van deze spreiding.

Statistics

Hoe vaak lieg je tegen je ouder(s)?

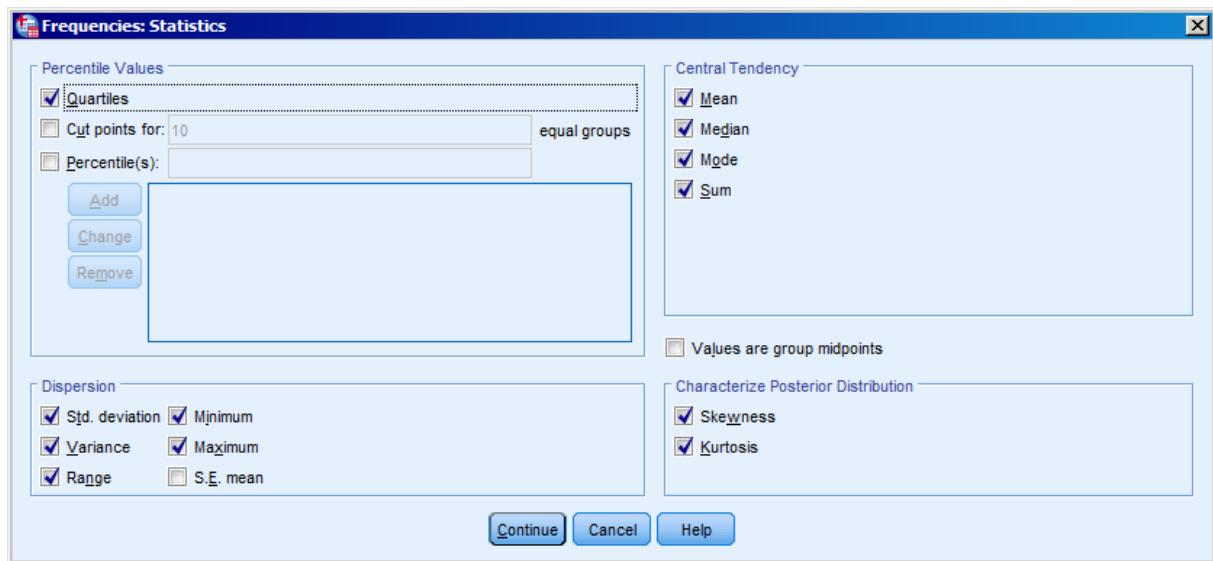
N	Valid	813
	Missing	2
Median		,00
Mode		0
Range		3
Minimum		0
Maximum		3
Percentiles	25	,00
	50	,00
	75	1,00

We kunnen dankzij de beschrijvende statistieken nog meer inhoudelijke informatie distilleren uit de ruwe data dan de informatie die we uit de frequentietabel kunnen opmaken. Uit de tabel met de beschrijvende statistieken kunnen we volgende parameters opvragen: de mediaan en de

modus, de variatiebreedte (range), de minimumwaarde, de maximumwaarde en de spreiding over de antwoordcategorieën op basis van de percentielen.

Parameters van centraliteit, spreiding en vorm voor metrische variabelen

Ga naar *Analyze Descriptive Statistics Frequencies*. We vragen alle univariate beschrijvende statistieken op van centraliteit (centrale tendensen), spreading (dispersion) en vorm (skewness en kurtosis).



We vroegen 4758 Belgische adolescenten naar hun bier/alcoholconsumptie van de afgelopen maand (*Denk terug aan de voorbije 30 dagen. Tijdens hoeveel verschillende gelegenheden heb je volgende zaken (bier of alcopops, wijn, sterke drank) gedronken ?*). Hieronder presenteren we de frequentietabel en de univariate beschrijvende statistieken. De variabele ‘bier/alcohol consumptie afgelopen maand’ is van het metrische niveau.



www.foksuik.nl

Bier/alcoholconsumptie afgelopen 30 dagen

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3.125	65,7	69,9	69,9
	1	472	9,9	10,6	80,5
	2	291	6,1	6,5	87,0
	3	188	4,0	4,2	91,2
	4	89	1,9	2,0	93,2
	5	110	2,3	2,5	95,6
	6	35	,7	,8	96,4
	7	24	,5	,5	97,0
	8	17	,4	,4	97,3
	9	10	,2	,2	97,6
	10	60	1,3	1,3	98,9
	12	7	,1	,2	99,1
	13	2	,0	,0	99,1
	14	1	,0	,0	99,1
	15	11	,2	,2	99,4
	16	2	,0	,0	99,4
	20	15	,3	,3	99,8
	25	2	,0	,0	99,8
	30	3	,1	,1	99,9
	40	1	,0	,0	99,9
	48	1	,0	,0	99,9
	50	3	,1	,1	100,0
	52	1	,0	,0	100,0
	Total	4.470	93,9	100,0	
Missing	>= 96	2	,0		
	ambiguous answer	27	,6		
	no answer	259	5,4		
	Total	288	6,1		
	Total	4.758	100,0		

4758 adolescenten werden gevraagd naar hun bier/alcoholconsumptie de afgelopen 30 dagen. Van deze 4758 hebben 288 jongeren (6.1%) de vraag niet of niet goed beantwoord. Dit betekent dat de rapportage van de resultaten gebaseerd zal zijn op de geldige antwoorden van 4470 adolescenten.

De overgrote meerderheid of 69.9% (3125) van de jongeren geeft geen enkele bier/alcoholconsumptie aan de afgelopen 30 dagen. 10.6% (472 respondenten) geeft aan tijdens 1 gelegenheid bier of alcohol te hebben gedronken in de afgelopen 30 dagen. 1.3% of 60 respondenten geven aan de afgelopen dertig dagen tijdens 5 gelegenheden bier of alcohol te hebben gedronken.

Statistics

Bier/alcoholconsumptie afgelopen maand

N	Valid	4.470
	Missing	288
Mean		1,07
Median		,00
Mode		0
Std. Deviation		3,040
Variance		9,244
Skewness		7,711
Std. Error of Skewness		,037
Kurtosis		93,982
Std. Error of Kurtosis		,073
Range		52
Minimum		0
Maximum		52
Percentiles	25	,00
	50	,00
	75	1,00

Wat kunnen we nu aflezen uit de tabel met de beschrijvende statistieken ?

N Valid (geldig): Aantal observaties (respondenten) waarvoor we een antwoord op de vraag hebben. 4470 respondenten hebben een ‘geldig’ antwoord gegeven op deze vraag.

Missing: aantal observaties waarvoor we geen antwoord op de vraag hebben. 288 respondenten hebben geen antwoord gegeven op deze vraag. Dit betekent dat $288/(4470 + 288)$ of 0.06 % van het totaal aantal bevraagde jongeren deze vraag niet beantwoord heeft. Noot: **alle berekeningen die de PC maakt, gebeuren op het aantal geldige observaties**

Percentiles: we krijgen hier de kwartieren te zien. De eerste 25 % van de observaties heeft een score 0 (geen bier/alcoholconsumptie de afgelopen 30 dagen), de middelste van de

gerangschikte observaties heeft een score van 0 en de respondenten die het derde kwartiel uitmaken, hebben een score van 1 (1 bier/alcoholconsumptie de afgelopen 30 dagen).

Parameters van centraliteit

Welke zijn nu de grote tendensen inzake bier/alcoholconsumptie gedurende de voorbije 30 dagen? Aangezien de variabele van het metrisch niveau is, kunnen we preciezere informatie verkrijgen dan het geval is voor variabelen gemeten op een nominaal of ordinaal meetniveau. De meest precieze informatie die we kunnen verkrijgen is het rekenkundig gemiddelde of mean. (We herhalen: het rekenkundig gemiddelde is de som van alle waarden gedeeld door het aantal respondenten)

Het **rekenkundig** gemiddelde bedraagt 1.07. Respondenten rapporteren een gemiddelde bier/alcoholconsumptie de afgelopen 30 dagen van 1.07.

De **mediaan** bedraagt .00. Als we de observaties zouden rangschikken op basis van het kenmerk ‘bier/alcoholconsumptie gedurende de afgelopen 30 dagen’ dan is de waarde die overeenkomt met de observatie die de groep netjes in twee gelijke delen verdeelt, .00. De mediaan ligt iets lager dan het gemiddelde.

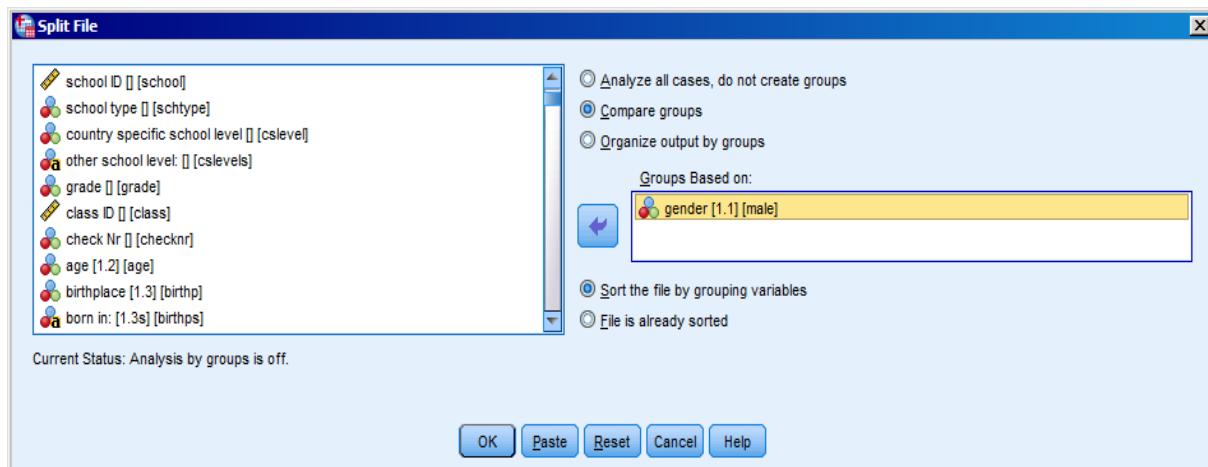
De **modus** bedraagt 0. De meest voorkomende waarde is 0. De meest typische jongere in deze steekproef heeft een bier/alcoholconsumptie gedurende de afgelopen 30 dagen, van 0.

Parameters van spreiding

Er zijn duidelijk grote verschillen inzake bier/alcoholconsumptie gedurende de afgelopen 30 dagen onder de respondenten in deze steekproef. De minimumwaarde bedraagt 0 en de maximumwaarde bedraagt 52. Daarom is de **variatiebreedte** 52. Dit wil zeggen dat bier/alcoholconsumptie gedurende de afgelopen 30 dagen varieert van 0 tot 52. Van een grote variabiliteit gesproken. Aangezien deze variabele van het metrische niveau is, kunnen we de **variantie** en **standaardafwijking** berekenen. Hieruit leiden we af hoe sterk onze respondenten van het gemiddelde verschillen. Is de variantie nul, dan is er geen spreiding en dan verschillen de respondenten niet van het gemiddelde. In dat heeft iedereen dezelfde waarde. Uiteraard kunnen we ook hier de kwartieren opvragen als we dat willen en op basis daarvan het verschil tussen de waarde die overeenkomt met het hoogste kwartiel en het laagste kwartiel bestuderen. De variantie bedraagt 9.244 en de standaardafwijking bedraagt 3.04. De respondenten verschillen dus van het rekenkundig gemiddelde.

Om de betekenis van de begrippen standaardafwijking en variantie ten volle te begrijpen, is het nog steeds zinvol om deze zelf te berekenen. Dit wordt in dit oefenboek behandeld in hoofdstuk II : univariate statistiek.

Je vraagt je wellicht af wat deze gegevens op zich zeggen. Op zich lijken sommige parameters van centraliteit en spreiding niet erg veel te zeggen. Ze vatten informatie samen uit een hele reeks van observaties over een bepaald kenmerk. Toch zijn deze voor beleidsmakers en onderzoekers belangrijk. Ongelijke spreiding roept veel criminologische vragen op. Waarom zijn er verschillen tussen individuen waar te nemen? Verder is het ook mogelijk dat er verschillen zijn tussen groepen. We presenteren hierna de beschrijvende statistieken van bier/alcoholconsumptie gedurende de afgelopen 30 dagen voor jongens en meisjes apart. We doen dit aan de hand van de bewerking *Split File*, waarna we de beschrijvende statistieken opvragen via *Analyze Descriptive Statistics Descriptives*.



Beschrijvende statistieken

geslacht		N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
female	Bier/alcoholconsumptie voorbije 30 dagen	2.245	40	0	40	,81	2,120	4,495
male	Bier/alcoholconsumptie voorbije 30 dagen	2.213	52	0	52	1,33	3,738	13,975

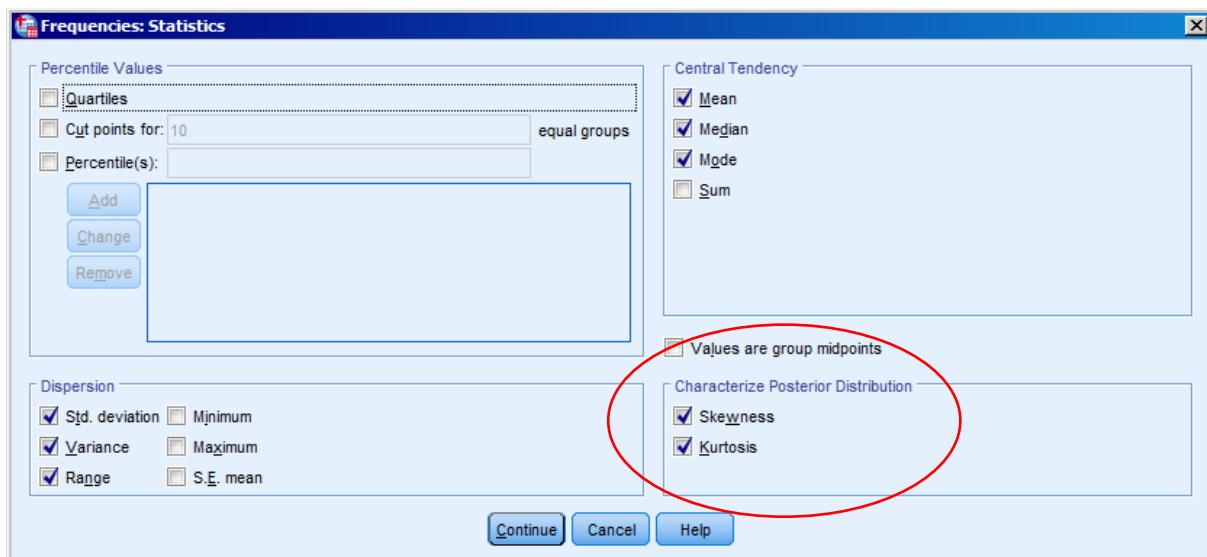
Hoe rapporteren we de resultaten ?

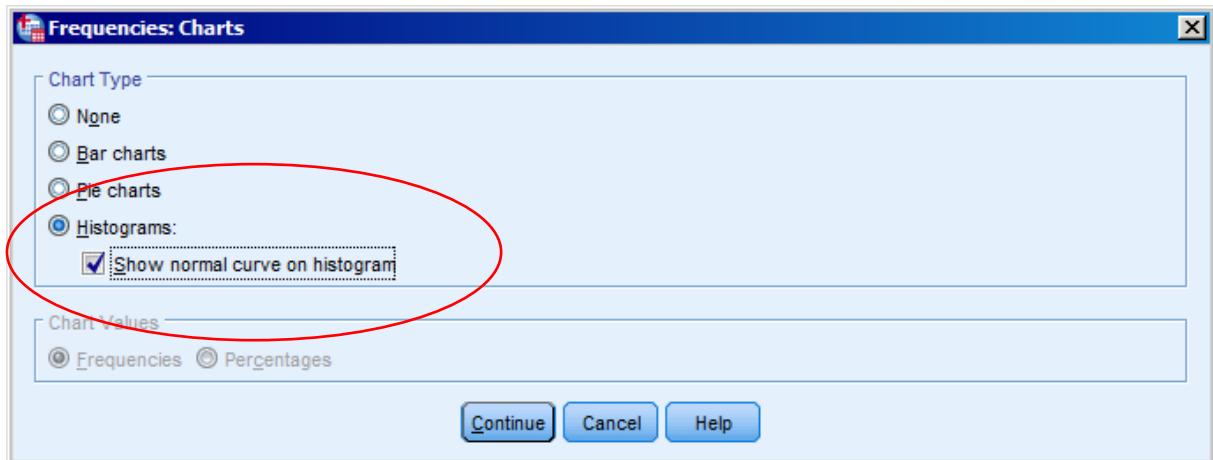
'Uit de beschrijvende statistieken kunnen we afleiden dat meisjes ($M=.81$) gemiddeld minder bier/alcoholconsumptie gedurende de laatste 30dagen rapporteren dan jongens ($M=1.33$). De standaardafwijking bij meisjes is lager dan bij jongens (respectievelijk: $SD: 2.120$ en $SD: 3.74$).

Bij jongens worden grotere verschillen waargenomen dan bij meisjes. De variatiebreedte bij meisjes bedraagt 40 en bij jongens 52. Beide groepen kenmerken zich door individuen die geen bier/alcoholconsumptie gedurende de laatste 30 dagen rapporteren, tot individuen die tot 40 verschillende gelegenheden met bier/alcoholconsumpties gedurende de afgelopen 30 dagen rapporteren bij meisjes en tot 52 bij de jongens.'

Parameters van vorm: skewness en kurtosis

De verdeling van kenmerken in de populatie kan verschillende vormen aannemen. De meest gekende verdeling is de symmetrische verdeling, ook wel normale verdeling genoemd. In het theoretische deel van deze syllabus wordt heel specifiek aandacht besteed aan de eigenschappen ervan. De verdeling van een kenmerk kan echter sterk afwijken van een normale verdeling. Er zijn verschillende manieren om inzicht te krijgen in de mate waarin een waargenomen verdeling van een variabele overeenkomt met een normale verdeling. We kunnen dit doen door parameters of grafieken op te vragen. Parameters die inzicht geven in de overeenkomst met een normale verdeling zijn scheefheid (skewness) en kurtosis.





We vragen een grafische voorstelling op van de variabele ‘bier/alcoholconsumptie gedurende afgelopen 30 dagen’ voor jongens en meisjes apart. We doen dit via *Analyze Descriptive Statistics Frequencies Charts*. Kies voor histogram (aangezien de variabele van het metrische niveau is) en vink aan *show normal curve on histogram*. Op die manier kunnen we visueel inspecteren in welke mate deze verdeling afwijkt van een normale verdeling. We vroegen ook de beschrijvende statistieken voor jongens en meisjes op met inbegrip van de parameters van vorm.

Statistics

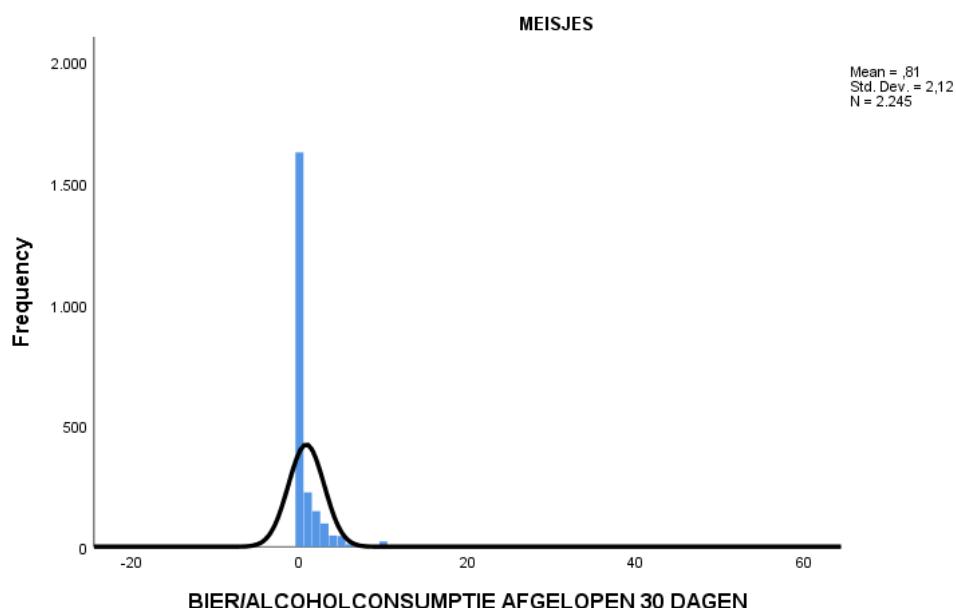
Bier/alcoholconsumptie afgelopen 30dagen

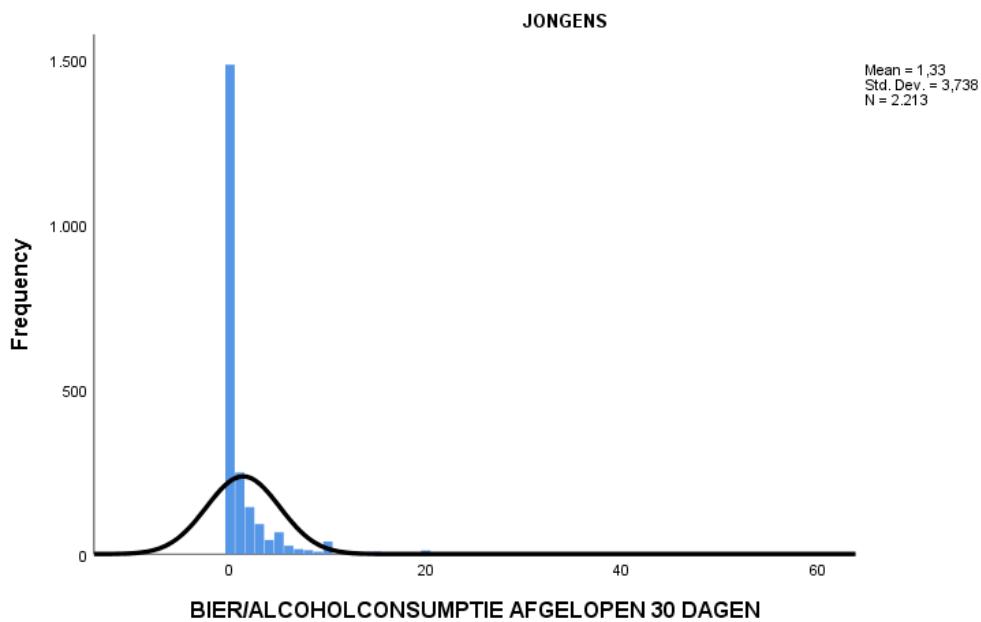
female	N	Valid	2.245
		Missing	125
		Mean	,81
		Median	,00
		Mode	0
		Std. Deviation	2,120
		Variance	4,495
		Skewness	6,266
		Std. Error of Skewness	,052
		Kurtosis	71,199
		Std. Error of Kurtosis	,103
		Range	40
male	N	Valid	2.213
		Missing	162
		Mean	1,33
		Median	,00
		Mode	0
		Std. Deviation	3,738
		Variance	13,975

Skewness	7,082
Std. Error of Skewness	,052
Kurtosis	73,355
Std. Error of Kurtosis	,104
Range	52

Onderstaande histogrammen tonen grafisch hoe het kenmerk ‘bier/alcoholconsumptie gedurende afgelopen 30 dagen’ verdeeld is onder jongens en meisje apart. Visuele voorstellingen zijn heel belangrijk. Ze tonen ons in een oogopslag hoe sterk een kenmerk varieert. Onmiddellijk valt op dat de meeste respondenten geen enkele bier/alcoholconsumptie gedurende de afgelopen 30 dagen rapporteren en dit geldt zowel voor jongens als voor meisjes. Daarentegen rapporteert een kleine groep herhaald bier/alcoholconsumptie en ook dit geldt voor beide geslachten. Hoe komt dit? Welke zijn nu de achtergrondkenmerken die dergelijke verschillen tussen individuen kunnen verklaren? Of zijn de verschillen eerder toevallig? Dit zijn vragen die sociaal wetenschappers interesseren.

Het fenomeen is statistisch niet normaal verdeeld. De verdeling neemt geen klokvorm aan. Hoe sterk wijkt deze verdeling dan af van de klokvorm? De staart van de verdeling kan langer zijn langs de rechterzijde, maar kan ook langer zijn langs de linkerzijde. Dit bestuderen we aan de hand van de scheefheid “skewness”. De verdeling kan ook platter zijn dan de eentoppige normaalverdeling. Dit bestuderen we aan de hand van de “kurtosis”.





Statistics

Bier/alcoholconsumptie afgelopen 30dagen

female	N	<u>Valid</u>	2.245
		Missing	125
		Skewness	6,266
		Std. Error of Skewness	,052
		Kurtosis	71,199
		Std. Error of Kurtosis	,103
male	N	<u>Valid</u>	2.213
		Missing	162
		Skewness	7,082
		Std. Error of Skewness	,052
		Kurtosis	73,355
		Std. Error of Kurtosis	,104

De skewness (scheefheid) geeft weer in welke mate een verdeling symmetrisch is. Bij de normale verdeling zijn modus, gemiddelde en mediaan gelijk aan elkaar en zijn de beide helften aan weerszijden van het gemiddelde elkaar spiegelbeeld en is de verdeling volledig symmetrisch. Als een variabele normaal verdeeld is, dan is de waarde van de skewness (scheefheid) en kurtosis gelijk aan nul. Positieve waarden voor de scheefheidsparameter en kurtosis parameter betekenen dat een bestudeerd kenmerk rechts schever en rechts platter is dan de situatie bij een normale verdeling, die overigens op het histogram werd aangebracht. We

krijgen hier de belangrijke informatie dat de verdeling van bier/alcoholconsumptie als kenmerk niet statistisch normaal verdeeld is, maar positief asymmetrisch is en dus een langere staart naar rechts heeft. Tevens is de verdeling rechts platter. De kurtosis (welving) is een maatstaf voor de relatieve platheid van de verdeling ten opzichte van de normale verdeling. De kurtosis is positiever als de top hoger ligt en negatiever als de top lager ligt. De normale verdeling heeft een kurtosis van nul. We laten de informatie over de “standard error of skewness” en “standard error of kurtosis” achterwege in dit hoofdstuk.

Zelf uitrekenen van de variantie en standaardafwijking in SPSS

Om de begrippen standaardafwijking en variantie ten volle te begrijpen, gaan we deze ontleden. De betekenis van deze abstracte begrippen vatten, kan nog altijd het best door deze eens zelf na te rekenen.

De variantie is de som van de gekwadrateerde afwijkingen van elke observatie tegenover het gemiddelde, gedeeld door $n-1$.¹ De variantie is dus de variatie gedeeld door $n-1$. De standaardafwijking is de vierkantswortel uit de variantie.

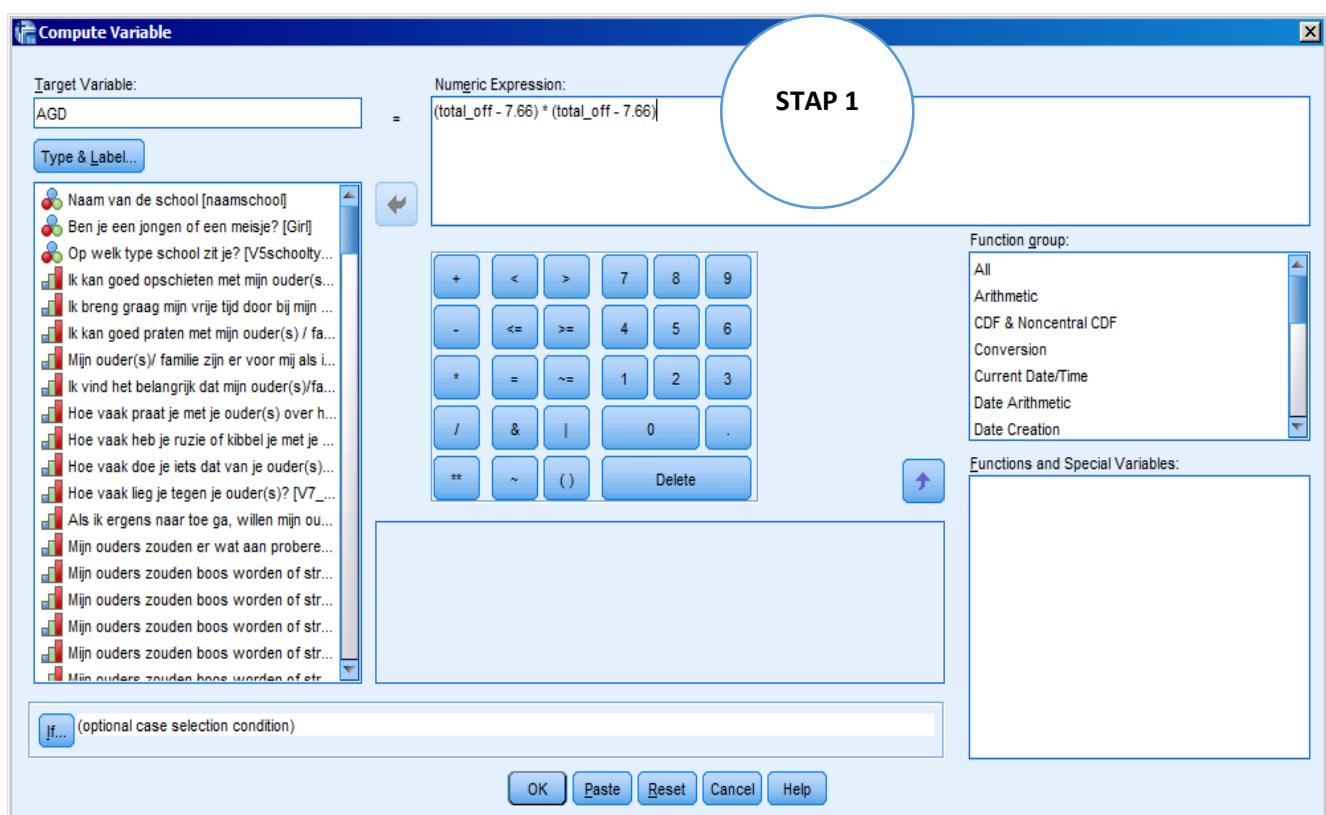
$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

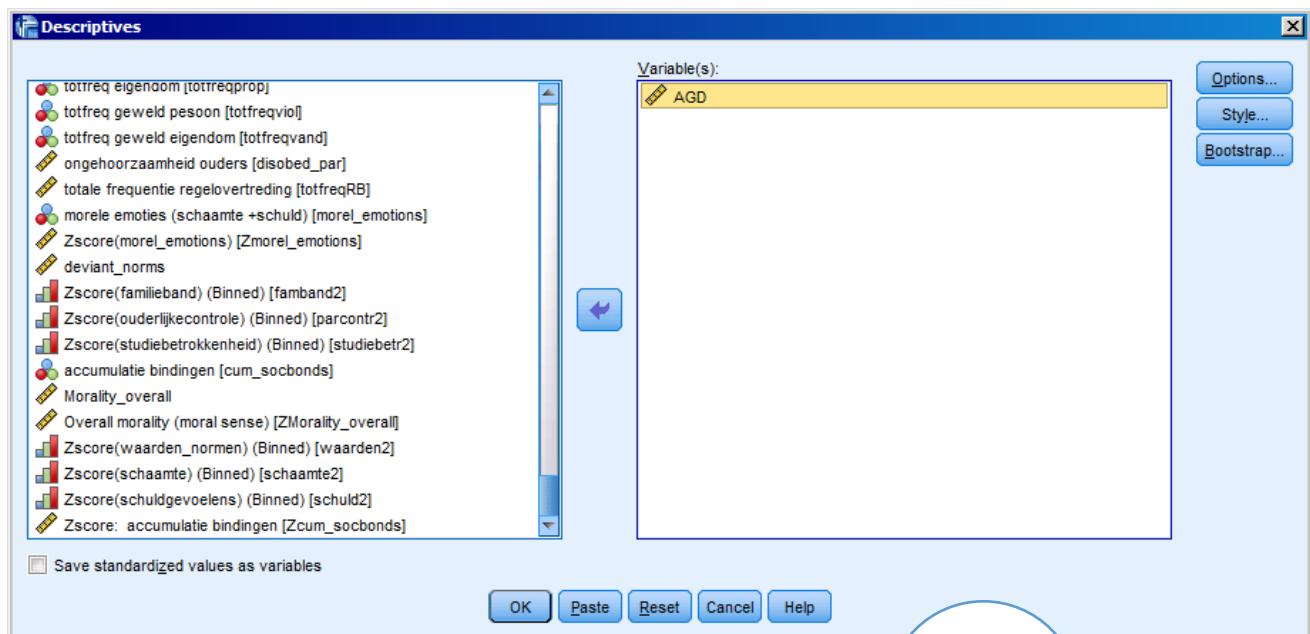
Teller: sum of squares (variatie)
Noemer: $n-1$

¹ We moeten wel kwadrateren, want de som van de afwijkingen tegenover het gemiddelde is steeds nul. Door het kwadraat van elke afwijking te nemen, krijgen we zicht op hoe groot de afwijking is, de richting speelt immers geen rol (hoger of lager dan het gemiddelde). We nemen $n-1$ als noemer wanneer de resultaten verkregen worden via een steekproef. Bij populatiegegevens is dit steeds n . In criminologisch onderzoek werken we bijna altijd met steekproeven.

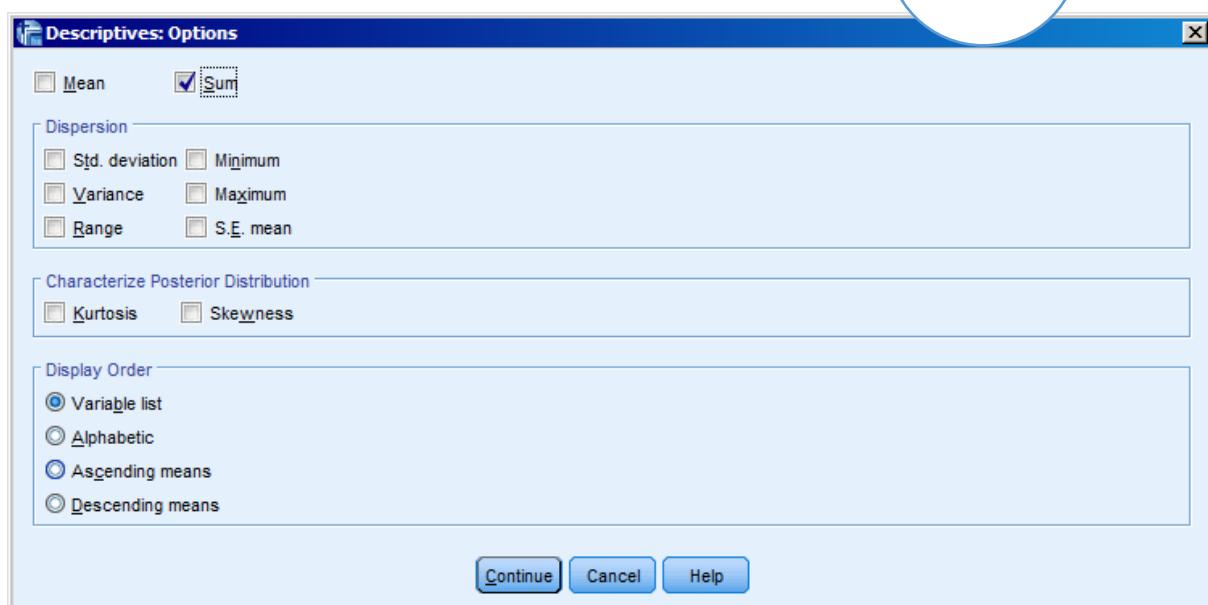
We vroegen aan 815 basisschoolkinderen hoeveel delicten zij al hadden gepleegd. 752 respondenten beantwoordden de vraag geldig. Het gemiddelde aantal gepleegde delicten bedroeg 7.66, de variantie 89.425 en de standaardafwijking 9.456. We kunnen zelf met SPSS narekenen of dit klopt, en dit in verschillende stappen:

- Het rekenkundig gemiddelde van de variabele ‘daderschap’ bedraagt: 7.66
- Laten we nu de afwijkingen van elke observatie tegenover het rekenkundig gemiddelde berekenen aan de hand van SPSS en deze vermenigvuldigen met zichzelf. We hoeven daarvoor een nieuwe variabele aan te maken die we AGD noemen (afwijking gemiddelde daderschap). We klikken op OK en we krijgen een nieuwe variabele (stap 1)
- Van deze variabele hebben we de som nodig volgens de definitie van de variantie: in SPSS doen we dit als volgt (stap 2)
- Het resultaat is de **sum of squares** of de variatie of de teller in de formule voor de berekening van de variantie (stap 3)





STAP 2

**Descriptive Statistics**

	N	Sum
AGD	752	67.157,89
Valid N (listwise)	752	

STAP 3

- ➔ Delen we de sum of squares door n-1, dan bekomen we: $67.157,89 / 751 = 89,425$
- ➔ De standaardafwijking is de vierkantswortel uit dit getal en dit bedraagt 9,456.
- ➔ Ter controle zie hieronder de beschrijvende statistieken van de variabele daderschap.

Statistics

daderschap

N	Valid	752
	Missing	63
Mean		7,66
Std. Deviation		9,456
Variance		89,425

6.2. Samenhang tussen twee variabelen

Om inzicht te krijgen in de samenhang tussen twee variabelen kunnen we in SPSS opvragen:

- **Kruistabel** voor de samenhang tussen nominale en ordinale variabelen (via opdracht *Analyze Descriptive Statistics Crosstabs*)
- **Spreidingsdiagram** of **scatterplot** waarbij de onafhankelijke variabele op de X-as wordt geplaatst en de afhankelijke op de Y-as. De waarnemingen worden gepresenteerd in de vorm van een puntenwolk. Spreidingsdiagrammen worden gemaakt via de opdracht *Graphs / Chart Builder* en vereist een minimaal ordinaal meetniveau. Spreidingsdiagrammen geven een goed inzicht in de mate waarin twee variabelen samenhangen en in de vorm van de samenhang (rechtlijnig of kromlijnig)

Samenhang tussen twee variabelen kan ook weergegeven worden in de vorm van een kengetal:

- **Chi-kwadraattoets:** om voor nominale en ordinale variabelen te bepalen of twee variabelen al dan niet onafhankelijk zijn van elkaar (*Analyze Descriptive Statistics Crosstabs Statistics*)
- **Phi en Cramers' V:** om voor nominale variabelen de mate van samenhang te bepalen (*Analyze Descriptive Statistics Crosstabs Statistics*)
- **Spearman's Rho en Kendall's Tau-b:** om voor ordinale variabelen de mate van samenhang te bepalen (*Analyze Descriptive Statistics Crosstabs Statistics*)
- **Pearson correlatiecoëfficiënt :** om voor metrische variabelen de mate van samenhang te bepalen (*Analyze Correlate Bivariate*)
- Partiële correlatie: als we willen controleren voor het effect van een derde variabele (*Analyze Correlate Partial*)

Zelf narekenen van de correlatiecoöefficiënt van Pearson aan de hand van SPSS (fictief voorbeeld)

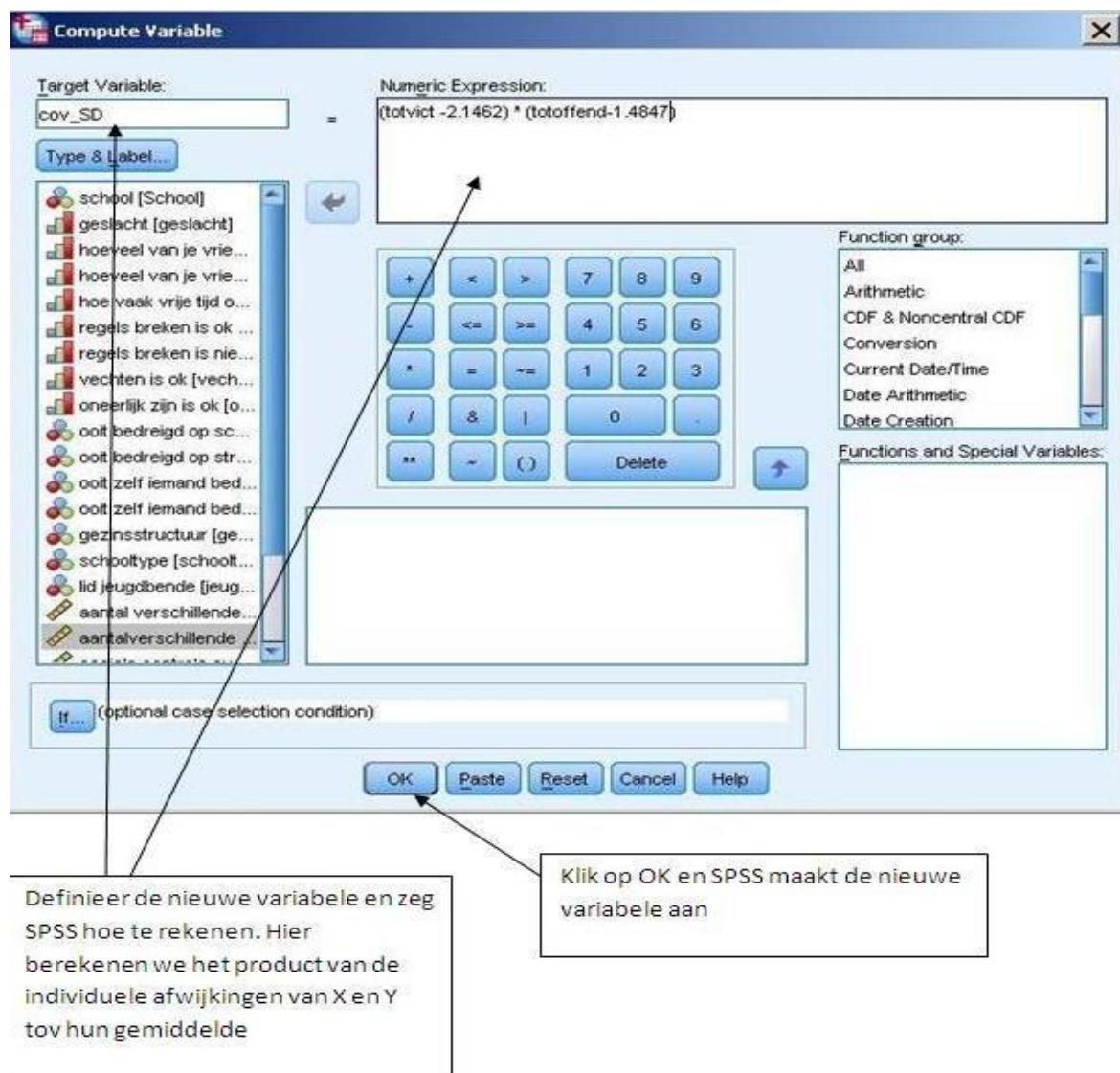
Covariatie of kruisproductensom of sum of squares (SSxy)	$SS_{xy} = \sum (X - \bar{X})(Y - \bar{Y})$
Covariantie (Sxy)	$S_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1}$
Correlatiecoöefficiënt van Pearson (Rxy)	$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$

In de teller van de formule van de correlatiecoöefficiënt staat de covariatie tussen X en Y. We beginnen dus met de covariatie te nemen tussen X en Y. Daartoe hebben we de gemiddelde waarden nodig van de beide variabelen. Gemiddelden, varianties en standaardafwijkingen verkrijgen we via de procedure ‘descriptives’.

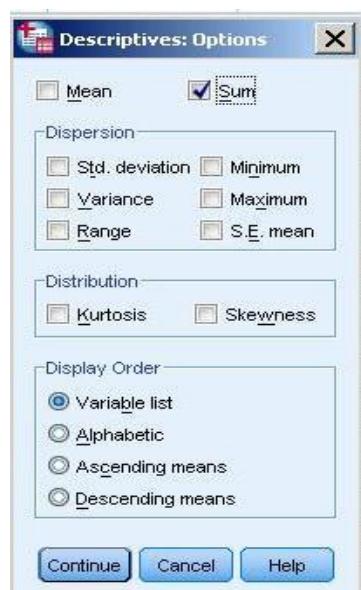
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
aantal verschillende delicten slachtoffer	1491	,00	10,00	2,1462	2,05203	4,211
aantal verschillende delicten gepleegd	1504	,00	13,00	1,4847	1,92052	3,688
Valid N (listwise)	1453					

Om de covariatie te berekenen doen we het volgende: we vermenigvuldigen de afwijkingen van beide variabelen tot hun respectievelijke gemiddelden. We gebruiken de procedure “compute”.



We moeten de som nemen van dit product. In SPSS doen we dat via “sum”



De som van de het product van de afwijkingen tov de respectievelijke gemiddelden is de volgende:

Descriptive Statistics

	N	Sum
cov_SD	1453	2532,45
Valid N (listwise)	1453	

Deze som is de kruisproductensom of covariatie, en moet eerst gedeeld worden door $(n-1)$ om tot de covariantie te komen.

De covariantie is dus $2532/1452 = 1.743$

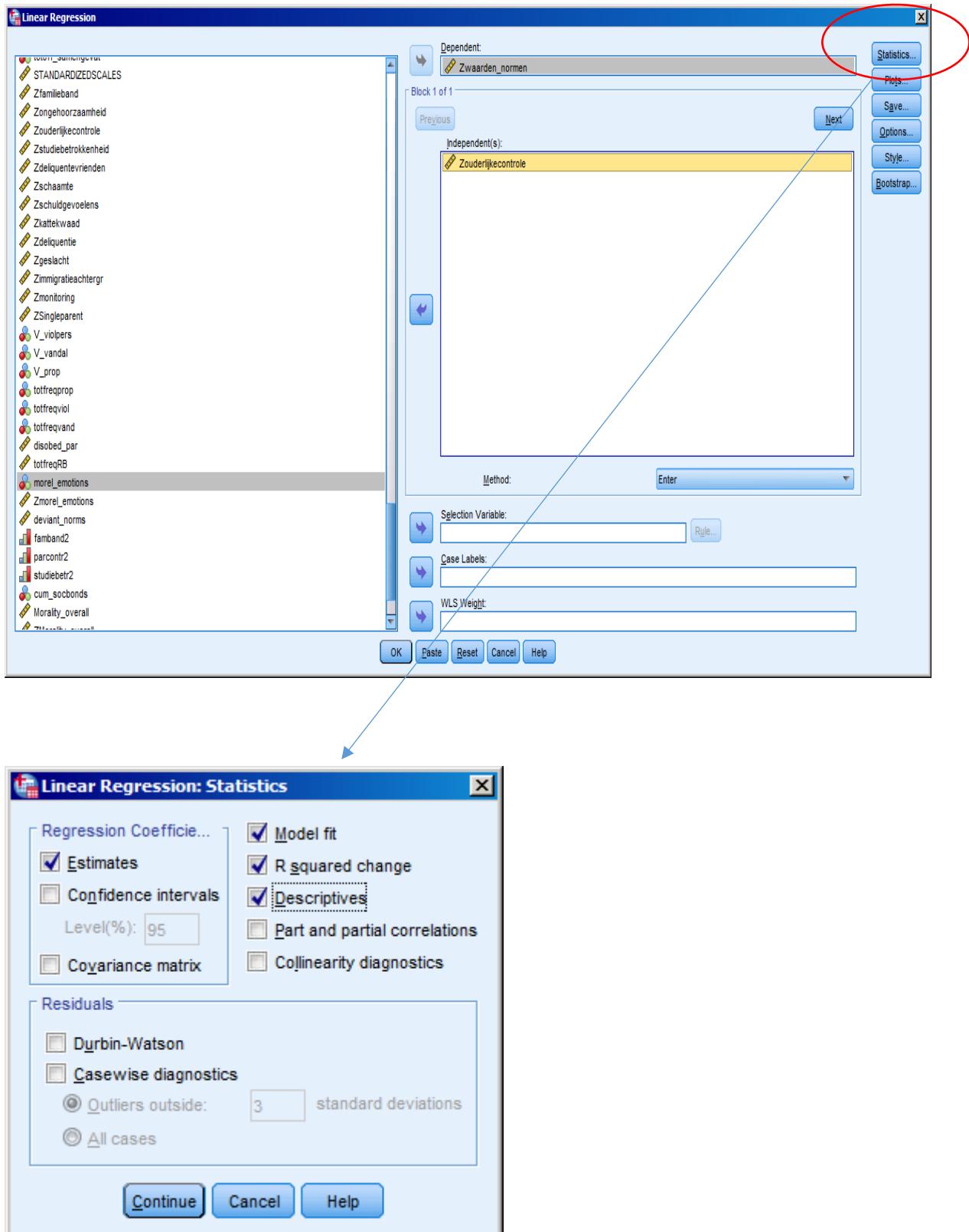
Deze waarden moeten we nog delen door het product van de standaardafwijkingen van de variabelen X en Y ofwel $(2.05 * 1.92) = 3.93 \rightarrow$ daarom is de correlatie $1.743 / 3.93 = 0.44$

Zoals je ziet, klopt onze eigen berekening met hetgeen SPSS doet! Met deze stapsgewijze naberekening hopen we dat je de filosofie achter een correlatie-analyse begrijpt. Hoewel we het meeste rekenwerk aan SPSS hebben overgelaten, hebben we je toch trachten te illustreren wat er eigenlijk gebeurt door een eenvoudige druk op de knop. Veel belangrijker is nu de vraag hoe we een correlatiecoëfficiënt inhoudelijk interpreteren vanuit een theoretische criminologische achtergrond. Vergeet niet dat je verwacht wordt als criminoloog een hypothese te kunnen formuleren over de aard van deze samenhang. De associatiemaat is symmetrisch. Maar wat denk je zelf over de relatie tussen daderschap en slachtofferschap? Beschouw je deze statistische relatie als uni-causaal? Zo ja, welke is oorzaak en gevolg? Of denk je eerder in termen van niet uni-causaal? Of gaat het hier om een schijneffect? Zo ja, wat zou dan de gemeenschappelijke oorzaak kunnen zijn? Wat denk je op basis van de criminologische theoretische bagage waarover je op dit moment beschikt?

6.3. Verklaren van een variabele op basis van één of meerdere onafhankelijke variabelen

Om te toetsen of er een lineair verband bestaat tussen een afhankelijke variabele en één of meerdere onafhankelijke variabelen wordt een regressie analyse gebruikt (respectievelijk enkelvoudige lineaire regressie of meervoudige of multiple lineaire regressie). De regressie analyse levert een vergelijking op waarmee de afhankelijke variabele numeriek kan worden verklaard. Belangrijke assumptie is dat er sprake is van een lineair (rechtlijnig) verband tussen metrische variabelen (interval of ratio-niveau). Een lineaire regressie analyse voeren we in SPSS uit via de opdracht *Analyze Regression Linear*.

The screenshot shows the IBM SPSS Statistics Data Editor window. The title bar reads "Scales Basisscholen Gent_LAATSTE VERSTE.sav [DataSet1] - IBM SPSS Statistics Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, and Help. The Analyze menu is currently open, displaying various statistical analysis options. The "Regression" option is expanded, and "Linear..." is highlighted with a yellow background. Below the menu, there is a large data grid containing 158 rows of variables. The columns include Name, Type, N, and M. The "Type" column shows mostly Numeric values, except for some categorical variables like "geanticipeerde schuld" which is Categorical. The "N" column shows sample sizes ranging from 2 to 19. The "M" column shows means ranging from 11 to 14. The data grid also includes columns for Label, Values, Missing, Columns, Align, Measure, and Role. At the bottom of the window, there are tabs for "Data View" and "Variable View", and a status bar at the bottom right indicating "IBM SPSS Statistics Processor is ready" and "Unicode:ON".



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	,386 ^a	,149	,148	,92471461	,149	129,326	1	739	,000

a. Predictors: (Constant), Zouderlijkecontrole Zscore(ouderlijkecontrole)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	110,587	1	110,587	129,326	,000 ^b
	Residual	631,917	739	,855		
	Total	742,503	740			

a. Dependent Variable: Zwaarden_normen Zscore(waarden_normen)

b. Predictors: (Constant), Zouderlijkecontrole Zscore(ouderlijkecontrole)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta	t		
1	(Constant)	-,005	,034		-,140		,889
	Zouderlijkecontrole	,386	,034	,386	11,372		,000
	Zscore(ouderlijkecontrole)						

a. Dependent Variable: Zwaarden_normen Zscore(waarden_normen)

'Uit de resultaten van de bivariate lineaire regressie-analyse blijkt dat 'ouderlijke controle' een matig sterk en positief effect heeft op 'morele normen en waarden' (β : .386). Dit effect is significant ($t=11.372$, $p < 0.001$). 14.8% van de variantie in 'morele normen en waarden' kan verklaard worden door 'ouderlijke controle' (R^2 : ,148). '