Beschrijvende statistiek

Beschrijvende statistiek

Het berekenen en interpreteren van tabellen en statistieken

Bregje van Groningen
Connie de Boer

Vierde druk

Boom

Woord vooraf

Voor veel studenten is statistiek een moeilijk vak. Daarom vinden we het belangrijk dat de uitleg zo duidelijk mogelijk is. Dat stelt hoge eisen aan de leesbaarheid. Ter wille van de leesbaarheid hebben we nauwelijks verwijzingen naar literatuur in de tekst opgenomen. In de literatuurlijst achter in het boek staan de publicaties waarop wij onze kennis hebben gebaseerd. Deze boeken gaan op sommige statistieken veel uitgebreider in. De literatuurlijst kun je daardoor ook beschouwen als aanbevolen literatuur voor een verdere verdieping van de kennis.

Wij denken dat een goed begrip van statistiek pas ontstaat als je weet hoe je statistische gegevens berekent. Daarom is de uitleg mede gebaseerd op de formules die je gebruikt voor een kengetal of statistische analyse. In de praktijk zul je statistieken zelden met de hand, op basis van formules, uitrekenen. Daarvoor bestaan heel geschikte rekenprogramma's, zoals SPSS. Voor het interpreteren van de resultaten is het echter van groot belang dat je weet wat de betekenis is van de statistische gegevens. Dan is het nuttig deze een paar keer zelf te hebben uitgerekend. Daarna kun je ook oefenen met behulp van een computerprogramma. In aparte kaders is uitgelegd hoe je de besproken analyses door SPSS kunt laten uitvoeren. Deze combinatie van uitleg van berekeningen met de hand en berekeningen met SPSS geeft je de mogelijkheid zelf sommen en oefeningen te maken waarbij je de berekeningen met de hand controleert aan de hand van de resultaten van SPSS.

Eerdere versies van dit boek zijn gebruikt in het onderwijs over onderzoeksmethoden en statistiek aan propedeusestudenten Communicatiewetenschap aan de Universiteit van Amsterdam. De reacties en het commentaar van de studenten en de docenten van deze cursussen hebben geleid tot vele verbeteringen van de tekst.

De volgende personen willen wij expliciet noemen als dank voor hun inhoudelijke en tekstuele bijdragen: Tiede Bijlsma, Sanneke Schouwstra, Rob de Lange, Wouter de Nooy, Mieke Sillekens, Reza Kartosen, Floris Müller, Johannes von Engelhardt, Rob Erven, Carel van Wijk, Nadine Bol, Jeroen Jonkman, Marianne Ouwehand, Gert van Driel en Rhianne Hoek. Verder hebben verschillende docenten en studenten van de cursus 'Methoden van Communicatie Onderzoek en Statistiek' commentaar geleverd op eerdere versies van deze tekst.

Nieuw in de vierde druk

In de vierde druk gaan we nog meer in op het belang van de operationalisatie, het kiezen van de juiste meetniveaus en de mogelijke consequenties van keuzes die gemaakt (moeten) worden in het analyseproces. Voorbeelden zijn aangepast aan (op het moment van schrijven) meer actuele zaken.

In hoofdstuk 3 is een aantal paragrafen toegevoegd waarin we onder andere ingaan op de normale verdeling als kansverdeling en *z*-scores bij het berekenen van kansen. Omdat het een boek over *beschrijvende* statistiek is, zullen we dat niet koppelen aan het toetsen van hypothesen en significantie.

Daarnaast is een nieuw hoofdstuk toegevoegd (hoofdstuk 4) waarin het ver- en bewerken van data in SPSS centraal staat, en de consequenties daarvan voor het meetniveau van de variabelen en de analyses die mogelijk zijn. In hoofdstuk 8 wordt uitgebreider dan in de derde druk ingegaan op het berekenen en interpreteren van een variantieanalyse. Hoofdstuk 9 en 10 zijn samengevoegd en uitgebreid. Het heet nu *Schaalconstructie*, en we bespreken daarin zowel de theoretische als statistische overwegingen bij het maken van een valide en betrouwbare schaal. In bijna alle hoofdstukken wordt meer stilgestaan bij de voorwaarden van bepaalde analyses en de consequenties wanneer niet aan de voorwaarden wordt voldaan.

Tot slot is de vierde druk aangepast aan de ontwikkelingen in de SPSS-software. Alle voorbeelden waarin SPSS wordt gebruikt, zijn nu uitgevoerd met SPSS 23.

Bregje van Groningen
Connie de Boer

Inhoud

Kaders

Inleiding

In dit boek staat de beschrijvende statistiek centraal. Maar wat is beschrijvende statistiek? Sociale wetenschappers hebben als doel kennis te genereren over de sociale werkelijkheid. Dat kan nieuwe kennis zijn, maar ook kennis waaruit blijkt dat wat we 'wisten' niet of niet helemaal klopt. Kennis kun je verkrijgen door het uitvoeren van een onderzoek. Wanneer je een onderzoek hebt uitgevoerd, wil je de resultaten van het onderzoek zo duidelijk mogelijk weergeven. Dat kan op een korte en overzichtelijke manier gebeuren door middel van kengetallen, tabellen of grafieken. Dit is precies waar het in de beschrijvende statistiek om gaat: het samenvattend beschrijven van de kenmerken van een groep onderzoekseenheden.
Een voorbeeld van een kengetal is het rekenkundig gemiddelde. Door de gemiddelde leeftijd van een groep uit te rekenen geef je een samenvattende beschrijving van een specifiek kenmerk van die groep, namelijk hun leeftijd. Je kunt dit kenmerk ook beschrijven door een tabel of een grafiek te gebruiken, zoals in tabel 1 is gedaan.

Tabel 1 Beschrijving van leeftijd in tabel

Leeftijd	Absolute frequentie	Percentage
18	5	25
19	6	30
20	1	5
21	2	10
22	2	10
24	4	20
Totaal	20	100

Om kengetallen, tabellen en grafieken te produceren moet een onderzoeker een aantal handelingen verrichten. Ten eerste moet hij de gegevens verzamelen en die verzamelde data onderbrengen in een datamatrix. Daarna moet hij de data verwerken (kengetallen bepalen en tabellen en grafieken maken). Deze gegevens vermeldt hij vervolgens in een onderzoeksverslag, waarbij het belangrijk is dat hij de statistieken op de juiste manier weergeeft en interpreteert. Deze fasen, de datamatrix, de data-analyse en de interpretatie van statistieken, worden in dit boek besproken.

In de komende hoofdstukken zal steeds over het algemeen eerst de theoretische achtergrond worden besproken. Het gaat daarbij om de vraag waarom je de analyses nodig hebt. Dit wordt geïllustreerd aan de hand van voorbeelden en waar mogelijk met een grafische weergave. De analyses zijn uitgevoerd met het computerprogramma SPSS, een programma dat in de sociale wetenschap veel wordt gebruikt. In deze vierde druk is gebruikgemaakt van SPSS 23 (Windows-versie) en Windows 10. Het is mogelijk dat de kaders en tabellen er wat betreft lay-out iets anders uitzien wanneer een andere versie van SPSS of Windows-versie, of een Mac is gebruikt. De inhoud is uiteraard steeds hetzelfde. We zullen in aparte kaders uitleggen hoe je met behulp van SPSS de besproken analyses zelf kunt uitvoeren.
In een wetenschappelijke tekst mag je nooit SPSS-tabellen opnemen, ook niet in de bijlagen. Je maakt als onderzoeker zelf tabellen waarin alleen de voor het onderzoek relevante gegevens staan. In dit boek wordt van deze regel afgeweken, omdat we willen laten zien hoe je een SPSS-output kunt lezen en interpreteren. In dit boek zijn de data, wanneer niet zelf verzonnen, afkomstig van databestanden die worden gebruikt door collega's van de afdeling Communicatiewetenschap, scripties van Masterstudenten, en enquêtes die zijn afgenomen door studenten van de module 'Methoden en Technieken voor Communicatieonderzoek'. Veel van deze data zijn bewerkt door de auteurs om de voorbeelden in het boek zo duidelijk mogelijk te maken.
Naast de interpretatie vanuit SPSS zal de handmatige berekening van verschillende analyses worden uitgelegd. Het zelf uitrekenen van statistieken op basis van de formules maakt duidelijk welke betekenis je eraan kunt toekennen. Dit zal het interpreteren van de gegenereerde kengetallen, tabellen en grafieken vergemakkelijken. Deze aspecten – nut en toepassing, voorbeelden, interpretatie en handmatige berekening – zullen in de komende hoofdstukken aan de orde komen bij de behandeling van de verschillende statistieken. Om zo dicht mogelijk in de buurt te komen van de waarden zoals ze in SPSS gepresenteerd zijn, rekenen wij altijd met drie decimalen achter de komma. In teksten over het interpreteren van de maten zullen we, zoals dat ook in de meeste wetenschappelijke artikelen wordt gedaan, ons beperken tot twee decimalen achter de komma.

In de beschrijvende statistiek gebruik je de gegevens van een groep onderzoekseenheden. Die groep kan een *steekproef* of een *populatie* zijn. De gebruikte statistieken zeggen dan iets over de samenstelling van die steekproef of over die populatie, afhankelijk van welke gegevens je gebruikt. Als het om een steekproef gaat, zegt de beschrijvende statistiek dus iets over de steekproef en niet over de populatie waaruit die steekproef afkomstig is. *Beschrijvende* statistiek onderscheidt zich van de *inferentiële* statistiek doordat je je bij de beschrijvende statistiek beperkt tot uitspraken over een groep onderzoekseenheden waarvan je de gegevens hebt. Wanneer je niet de hele populatie hebt onderzocht, kun je over de populatie niets met zekerheid zeggen. Op basis van de steekproefgegevens kun je echter wel met een bepaalde mate van waarschijnlijkheid uitspraken

doen over de populatie. Dit gebeurt in de *inferentiële statistiek*. Op basis van steekproefgegevens maak je dan een schatting van een populatiewaarde.

De statistieken of kengetallen uit de beschrijvende en de inferentiële statistiek geef je aan met symbolen, met letters. Zo duid je het gemiddelde in de populatie aan met een μ (mu), en een gemiddelde in een steekproef met een $\bar{x}$ (*x* streep) of in een wetenschappelijk artikel met *M* (mean). De standaarddeviatie in een populatie geef je aan met een σ (sigma). Als je de standaarddeviatie van de steekproef (*s*) gebruikt als schatter van de standaarddeviatie in de populatie, ben je met inferentiële statistiek bezig. De formules van σ (standaarddeviatie van de populatie) en *s* (standaarddeviatie van de steekproef) vertonen een klein verschil. Bij σ wordt door *n* (aantal onderzoekseenheden) gedeeld, en bij *s* door *n – 1* (aantal vrijheidsgraden[1]). Als je door *n* zou delen, dan schat je de standaarddeviatie in de populatie iets te laag. Ditzelfde geldt niet alleen voor de standaarddeviatie, maar ook voor andere kengetallen die in dit boek worden behandeld.
Bij de behandeling van de beschrijvende statistiek zou het logisch zijn om te kiezen voor de kengetallen waarbij door *n* wordt gedeeld. De statistieken worden immers niet gebruikt als schatter voor de populatieparameters. Hier hebben we niet voor gekozen, omdat we naast het met de hand uitrekenen van statistische gegevens aandacht besteden aan SPSS. Je kunt de met de hand uitgerekende statistieken controleren door de gegevens in een SPSS-bestand in te voeren. Het is ook mogelijk om op basis van een SPSS-bestand zelf oefeningen te maken voor het met de hand uitrekenen van statistieken. Aangezien je SPSS ook gebruikt voor de inferentiële statistiek, zou er – zeker als het gaat om een klein aantal onderzoekseenheden – een verschil kunnen bestaan tussen de output van SPSS en de zelf uitgerekende gegevens als je de formules uit de beschrijvende statistiek gebruikt. Daarom hebben we ervoor gekozen de formules te gebruiken die horen bij de inferentiële statistiek.

In hoofdstuk 1 staan de basiselementen van de beschrijvende statistiek centraal. De centrummaten (modus, mediaan en gemiddelde) en spreidingsmaten worden respectievelijk in hoofdstuk 2 en 3 beschreven. In hoofdstuk 3 wordt ook de normale verdeling als kansverdeling besproken. Hoofdstuk 4 geeft informatie over hoe je in SPSS je data kunt bewerken. De hoofdstukken 5, 6, 8 en 9 gaan over de associatiematen op nominaal, ordinaal en interval- en rationiveau. In hoofdstuk 7 wordt aandacht besteed aan tabelsplitsing. Hoofdstuk 10 gaat over schaalconstructie en behandelt factoranalyse en betrouwbaarheidsanalyses.

Website

Bij dit boek hoort een website met aanvullend materiaal: www.beschrijvendestatistiek.nl. Op deze website staan per hoofdstuk opdrachten. Studenten kunnen zo oefenen met de lesstof van het hoofdstuk. Op pagina 4 van dit boek staat een persoonlijke inlogcode voor toegang tot de website.

Noot

1 Dit is het aantal keren dat een waarde 'vrijelijk kan variëren'. Als je van een groep mensen de gemiddelde leeftijd weet, en je de leeftijden weet van alle individuen uit die groep op één na ($n-1$), staat de leeftijd van die laatste persoon vast. Die kan niet meer variëren. De leeftijd van deze laatste persoon kun je precies uitrekenen op basis van de leeftijden van de anderen uit de groep en de gemiddelde leeftijd.

Basiselementen 1

Je doet onderzoek om iets over de werkelijkheid te weten te komen. Op basis van dat onderzoek kun je dan uitspraken doen over de werkelijkheid. Daarbij moet duidelijk worden over wie of wat je op basis van het onderzoek een uitspraak doet. Dat zijn de objecten of *onderzoekseenheden,* de personen of zaken over wie je iets zegt.
Als je op basis van een onderzoek bijvoorbeeld de conclusie trekt dat de gemiddelde leeftijd van de eerstejaarsstudenten van een universiteit 19,7 jaar is, zeg je op basis van dit onderzoek iets over eerstejaarsstudenten. Dat zijn dan de onderzoekseenheden. Van deze studenten beschrijf je een kenmerk, namelijk de leeftijd. In het onderzoek kun je meer kenmerken van de studenten hebben verzameld, zoals geslacht, vooropleiding en studiekeuze. Deze kenmerken (leeftijd, geslacht, vooropleiding en studiekeuze) zijn de *variabelen* in het genoemde onderzoek.
Onderzoekseenheden hoeven niet altijd personen te zijn. Je kunt op basis van een onderzoek ook uitspraken doen over de lengte van voorpagina-artikelen in dagbladen. In dat geval zijn de voorpagina-artikelen de onderzoekseenheden, de objecten waarover je een uitspraak doet. De lengte is hier een kenmerk van de onderzochte artikelen en is daarom een variabele in het onderzoek. Met behulp van statistiek kun je precieze uitspraken doen over de kenmerken van onderzoekseenheden, zoals 'de gemiddelde leeftijd van eerstejaarsstudenten is 19,7 jaar' en 'de meeste voorpagina-artikelen zijn korter dan twee kolommen'.
Een ander voorbeeld: 'televisiekijkers met een hoge opleiding kijken vaker naar het nieuws dan televisiekijkers met een lage opleiding'. In deze uitspraak wordt iets gezegd over televisiekijkers. In het onderzoek zijn televisiekijkers de onderzoekseenheden. In het voorbeeld zijn de kenmerken van die televisiekijkers: opleiding en de frequentie waarmee naar het nieuws wordt gekeken. 'Opleiding' en 'frequentie nieuws kijken' zijn de variabelen in het onderzoek. Meer voorbeelden van onderzoekseenheden en variabelen staan in paragraaf 1.4.

1.1 Datamatrix

Variabelen, de kenmerken van onderzoekseenheden, kunnen verschillende waarden hebben. Bij sommige kenmerken zijn de waarden al een getal, bij andere kenmerken zou je voor de voorkomende categorieën een getal kunnen verzinnen. De waarden van bijvoorbeeld het kenmerk leeftijd zijn getallen die direct gerelateerd zijn aan de werkelijkheid. Als een persoon 21 jaar oud is, is

het logisch dat deze persoon de waarde 21 krijgt voor de variabele 'leeftijd in jaren'. Maar bijvoorbeeld de variabele geslacht heeft geen vaststaande numerieke waarde. Om in de statistiek toch op een geordende wijze iets te kunnen zeggen over de onderzoekseenheden, krijgen de categorieën 'man' en 'vrouw' waarin de variabele 'geslacht' kan worden onderverdeeld, wel een numerieke waarde om de dataverwerking te vergemakkelijken. Je zou kunnen besluiten vrouwen de waarde 1 te geven en mannen de waarde 2. Op die manier kun je alle onderzoekseenheden voorzien van een numerieke waarde voor het kenmerk 'geslacht'. Al deze kenmerken van onderzoekseenheden kun je onderbrengen in een *datamatrix*. Een datamatrix is een spreadsheet waarin per onderzoekseenheid alle kenmerken als afzonderlijke variabelen worden beschreven. De onderzoekseenheden staan in de rijen van de datamatrix en de variabelen in de kolommen (tabel 1.1).
Stel dat je lekker op een terrasje zit met je vrienden. Je vraagt je af of je vrienden naar dezelfde televisiezenders kijken als jij. Om dat uit te vinden, maak je een lijstje met een aantal televisiezenders waarnaar je zelf regelmatig kijkt en gaat iedereen vragen hoeveel uur zij ongeveer per week naar die zender kijken. Daarbij schrijf je ook op welke leeftijd die persoon heeft en of het een man of een vrouw is. Omdat je alles in getallen wilt uitdrukken, stel je dat als iemand vrouw is zij de waarde 1 krijgt, en als iemand man is de waarde 2 (we hadden ook vrouw de waarde 2 kunnen geven en man de waarde 1, of een symbool kunnen toevoegen, waarbij vrouw = ♀ en man = ♂). De datamatrix ziet er dan als volgt uit.

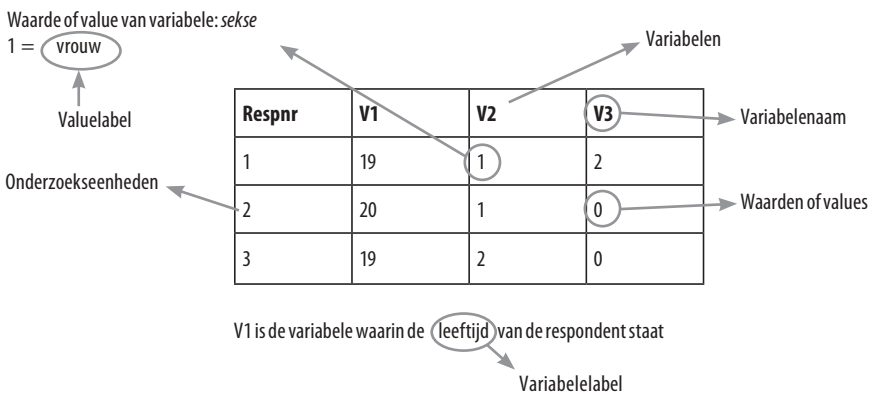
Tabel 1.1 Voorbeeld van een datamatrix

Persoon	Leeftijd	Sekse	NPO1	NPO3	RTL4	RTL5	Net5	BBC
1	19	1	2	0	2	2	4	0
2	20	1	0	0	1	2	2	1
3	19	2	0	1	1	1	0	3
4	22	1	3	2	2	3	3	2
5	24	2	1	0	3	1	1	1
6	21	2	0	3	2	0	1	1
7	20	1	2	2	1	2	3	2

In de eerste rij staan per kolom de namen van de variabelen die je in je onderzoek hebt gemeten. In de cellen daaronder staan de waarden die de respectievelijke onderzoekseenheden hebben op die variabelen. Persoon 1 is dus een 19-jarige vrouw, die twee uur naar NPO1, RTL4 en RTL5 kijkt, vier uur naar Net5, en niet naar NPO3 en de BBC.
Voor elke afzonderlijke variabele kun je een *frequentieverdeling* maken om uitspraken te doen over de percentuele verdeling van de onderzoekseenheden over de waarden van die variabele. In dit voorbeeld, waarbij we ons hebben beperkt tot zeven onderzoekseenheden, is 42,9% man (drie van de zeven personen) en

57,1% vrouw (vier van de zeven personen). Voor elke variabele (elke kolom in de datamatrix) kun je een dergelijke frequentieverdeling maken.

Om met de terminologie wat vertrouwd te raken laten we nogmaals een klein gedeelte van de datamatrix zien, en zullen we stilstaan bij de verschillende begrippen die hierbij horen. We hadden al gezien dat in de rijen de onderzoekseenheden staan, en dat deze variëren op een aantal kenmerken (vandaar ook de naam: variabelen). In de datamatrix behorende bij tabel 1.1 is duidelijk te zien wat er met de variabelen bedoeld wordt. Met 'leeftijd' wordt de leeftijd van de respondent bedoeld, en met 'sekse' het geslacht van de respondent. Met 'NPO1' kunnen echter verschillende dingen worden bedoeld. Is de onderzoekseenheden gevraagd hoeveel uur ze per week kijken? Of per dag? Of misschien is hun wel een schaal voorgelegd waarop ze konden antwoorden van 0 = nooit tot 5 = heel vaak.
De namen die boven de kolommen van een datamatrix staan, zijn de *variabelennamen*. We zouden ook voor heel andere variabelennamen kunnen kiezen, zoals voor V1, V2 en V3, omdat deze informatie bijvoorbeeld correspondeert met respectievelijk vraag 1, vraag 2 en vraag 3 in een vragenlijst (zie figuur 1.1).



Figuur 1.1 Terminologie bij datamatrix

V3 is hier dus een variabelennaam. Wat je dan feitelijk bedoelt met die variabelennaam, maak je kenbaar in het *variabelenlabel*. Het label van V1 is in voorstaand voorbeeld dus leeftijd in jaren. Het label van V2 is sekse, en van V3 is het label aantal uur dat naar NPO1 wordt gekeken. De getallen die in de matrix staan, noemen we *waarden* of, in het Engels, *values*. De waarden die bij de variabele V1 horen (leeftijd), zijn gemakkelijk te interpreteren: 19 betekent dat deze respondent 19 jaar oud is. De waarde 1 van de variabele V2 is al moeilijker te interpreteren. Daarom worden ook de waarden voorzien van een label: het *valuelabel*. Daarin wordt aangegeven wat bedoeld wordt met de waarden. In ons voorbeeld is het valuelabel van de waarde 1 van de variabele V2: vrouw. Het valuelabel van de waarde 2 van de variabele V2 is hier man. Om te begrijpen wat er met de verschillende variabelennamen, variabelenlabels en valuelabels

wordt bedoeld, is het nodig om deze informatie ergens te vermelden. Dit kan bijvoorbeeld in een codeboek. In SPSS kun je deze informatie zelf toevoegen in het tabblad 'Variable View' (zie kader 1.1).

1.2 Frequentietabellen

Een variabele (kenmerk) met de daarbij behorende waarden kun je op een overzichtelijke manier presenteren in een *frequentietabel*. Stel dat je in de zomer weer met wat vrienden op een terrasje zit en jij bent aangewezen om de drankjes te gaan halen. Je kunt proberen alles te onthouden, maar op een bierviltje de drankjes turven is gemakkelijker. Door te turven maak je een overzicht van het aantal keer dat een waarde voorkomt. Het tellen van de streepjes brengt je op de *absolute frequentie*.

Tabel 1.2 Turven en tellen van drankjes

Drankje	Aantal (geturfd)	Absolute frequentie	Percentage
Bier	𝍸 |||	8	47,1
Rosé	𝍸	5	29,4
Cola Light	|	1	5,9
Cappuccino	|||	3	17,6
Totaal		17	100

Dit is de basis voor het opstellen van een frequentietabel. Uit tabel 1.2 blijkt dat van de zeventien mensen acht een biertje willen, vijf een rosé enzovoort.

Een absolute frequentie kan moeilijk te interpreteren zijn, zeker wanneer je meerdere frequentieverdelingen met elkaar wilt vergelijken. Daarom is het handig om naast de absolute frequenties percentages te geven. De percentages bereken je door de absolute frequentie waarmee een specifieke waarde voorkomt te delen door het totaal aantal eenheden. In vorenstaand voorbeeld heeft $\frac{8}{17} = 0,471 = 47,1\%$ van je vrienden een biertje besteld.

De week daarop zit je weer op een terras, maar nu met 22 vrienden. Het aantal bier, rosé en cola light is hetzelfde, maar in plaats van drie, worden nu acht cappuccino's besteld.

Absoluut gezien worden dezelfde hoeveelheden bier, rosé en cola light besteld, maar relatief gezien (kijkend naar de percentages, rekening houdend met het totale aantal drankjes) wordt er minder bier, rosé en cola light besteld.

Uit tabel 1.3 blijkt dat nu 36,4% een biertje bestelt: $\frac{8}{22} = 0,364$.

Tabel 1.3 Aantal drankjes (absoluut en in percentages)

Drankje	Aantal (geturfd)	Absolute frequentie	Percentage
Bier	𝍸 |||	8	36,4
Rosé	𝍸	5	22,7
Cola Light	|	1	4,5
Cappuccino	𝍸 |||	8	36,4
Totaal		22	100

1.2.1 Hoe ziet dit eruit in SPSS?

Stel, je wilt van de zeventien vrienden tijdens het eerste terrasbezoek weten hoe oud ze zijn. Je pakt weer je bierviltje, vraagt ieders leeftijd en gaat turven. Later die dag voer je je gegevens in SPSS in, en je laat SPSS een frequentietabel maken (zie kader 1.1). Zoals is te zien in tabel 1.4, geeft SPSS naast de absolute frequentie (*Frequency*) en het percentage (*Percent*) ook het geldige percentage (*Valid Percent*) en het cumulatieve percentage (*Cumulative Percent*).

Tabel 1.4 Frequentieverdeling van de variabele leeftijd (SPSS-output)

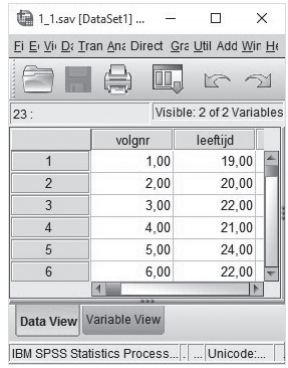
Leeftijd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19	4	23,5	23,5	23,5
	20	5	29,4	29,4	52,9
	21	3	17,6	17,6	70,6
	22	4	23,5	23,5	94,1
	24	1	5,9	5,9	100,0
	Total	17	100,0	100,0	

In de eerste kolom (*Valid*) zie je de waarden die de variabele leeftijd in ons voorbeeld heeft. Je vrienden zijn 19, 20, 21, 22 of 24 jaar oud. In de tweede kolom *(Frequency)* staan de absolute frequenties, het aantal keer dat een bepaalde waarde voorkomt. In dit geval zijn de percentages in de vierde kolom *(Valid Percent)* identiek aan de percentages in de derde kolom *(Percent).* In paragraaf 1.2.2 zullen we bespreken in welke situaties dit niet het geval is. In deze kolommen kunnen we aflezen dat 23,5% van de onderzoekseenheden 22 jaar is. In de kolom *Cumulative Percent* worden de percentages van elke volgende waarde bij de voorgaande opgeteld (23,5 + 29,4 = 52,9 enzovoort). Je zou aan de hand van deze kolom kunnen concluderen dat 52,9% van de onderzoekseenheden (in dit geval je vrienden op het terrasje) 20 jaar of jonger is.

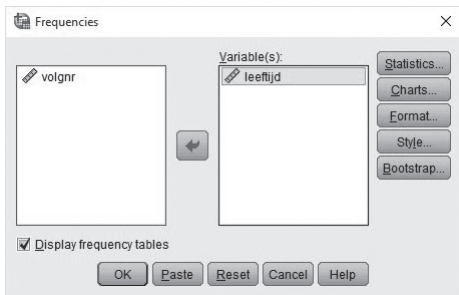
SPSS Invoeren van gegevens en het maken van een frequentietabel

Het invoeren van gegevens gebeurt in de Data View van SPSS. Per onderzoekseenheid kun je bijvoorbeeld na een uniek volgnummer voor elke onderzoekseenheid de waarde voor de variabele leeftijd intikken (zie figuur A). De namen van de variabelen (hier: 'volgnr' en 'leeftijd') kun je invoeren op de pagina die achter deze datamatrix ligt (klik linksonder op Variable View).



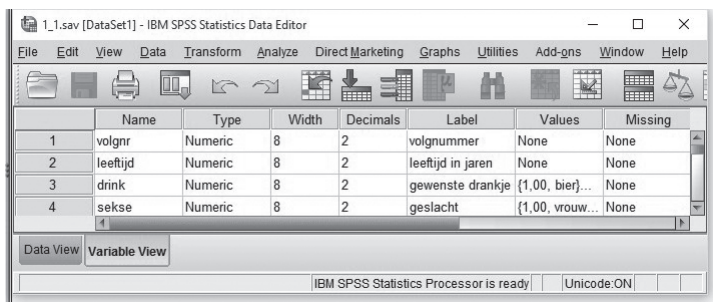
Figuur A Datamatrix

Als de data in SPSS zijn ingevoerd, kun je om een frequentietabel vragen. Ga via Analyze → Descriptive Statistics → Frequencies om vervolgens in het Frequencies-venster (zie figuur B) de variabelen te selecteren waar je een frequentietabel van wilt. SPSS maakt zelf het outputbestand waarin je deze tabel kunt vinden.



Figuur B Frequencies-venster

Door in SPSS links onderin op 'Variable View' te klikken, krijg je een overzicht van jouw variabelen te zien.



Figuur C Variable View

1.2.2 Missing values

Je vrienden willen je best vertellen hoe oud ze zijn. Wanneer je echter willekeurig mensen op een terras gaat vragen hoe oud ze zijn, kan het gebeuren dat ze je dat niet willen vertellen. Dan heb je voor die personen (onderzoekseenheden) dus geen informatie over het kenmerk leeftijd. Deze ontbrekende waarden noem je *missing values*. Je neemt deze mensen wel mee in je onderzoek naar de kenmerken van de personen op een terras. Maar als je wilt weten met welk percentage elke leeftijd vertegenwoordigd is, wil je soms niet dat deze onderzoekseenheden met onbekende leeftijden meetellen bij de berekening van de percentages. In tabel 1.5 is te zien hoe dit er in SPSS uitziet.

Tabel 1.5 Frequentieverdeling naar leeftijd met missing values (SPSS-output)

Leeftijd

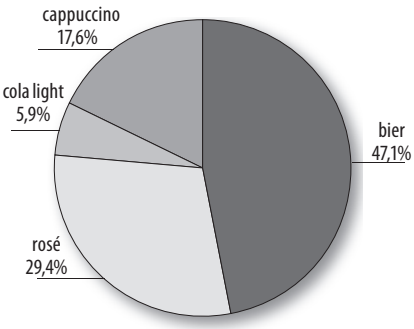
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19	4	20,0	23,5	23,5
	20	5	25,0	29,4	52,9
	21	3	15,0	17,6	70,6
	22	4	20,0	23,5	94,1
	24	1	5,0	5,9	100,0
	Total	17	85,0	100,0	
Missing	geen antwoord	3	15,0		
Total		20	100,0		

In totaal zitten twintig personen op het terras. Daarvan hebben drie mensen geen antwoord willen geven op de vraag naar hun leeftijd. Er is nu een verschil tussen Percent en Valid Percent. Bij Percent worden deze mensen namelijk wel meegerekend (15% van de ondervraagden heeft geen antwoord gegeven). Het percentage 19-jarigen van alle mensen op het terras is 20%. Of dat een zinnig percentage is, is nog maar de vraag, want de drie die geen antwoord hebben gegeven zouden ook 19 jaar kunnen zijn, maar informatie daarover ontbreekt. Daarom kun je in dit geval beter kijken in de kolom met Valid Percent, het geldige percentage. Hierin worden de mensen die geen antwoord hebben gegeven niet meegerekend. Dan kun je constateren dat 23,5% van de mensen op het terras die deze vraag hebben beantwoord 19 jaar is. In paragraaf 4.2 wordt verder ingegaan op hoe je in SPSS waarden missing kunt maken en wat daar de consequenties van kunnen zijn in je onderzoek.

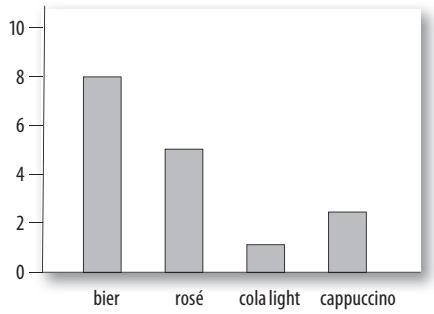
1.2.3 Grafieken

Een frequentietabel kun je ook grafisch weergeven. Een grafiek geeft geen extra informatie, maar kan visueel snel duidelijk maken wat de verdeling is van de waarden van een variabele. Bij een frequentieverdeling is het mogelijk om een

taart-, cirkel- of staafdiagram te gebruiken. In figuren 1.2 en 1.3 wordt visueel duidelijk gemaakt welke dranken door veel en welke dranken door weinig van je vrienden zijn besteld. De diagrammen geven de verdeling van de percentages of de frequenties weer.



Figuur 1.2 Taartdiagram drankjes

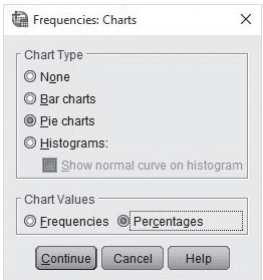


Figuur 1.3 Staafdiagram drankjes

In het taartdiagram kun je zien dat de meeste mensen een biertje hebben besteld (47,1%) en dat cola light door het kleinste aantal mensen is besteld (5,9%). Ditzelfde is af te lezen in het staafdiagram: acht mensen bestelden een biertje, en één persoon een cola. Een taart- of cirkeldiagram wordt meestal gebruikt om aan te geven wat relatief vaak voorkomt (percentages), terwijl staafdiagrammen meestal worden gebruikt om absolute aantallen weer te geven. Beide worden alleen gebruikt wanneer een variabele niet zo heel veel waarden heeft. Wanneer je in een enquête aan 200 onderzoekseenheden naar hun leeftijd vraagt en daar vijftig verschillende leeftijden uitkomen, is het niet erg overzichtelijk om daar een cirkeldiagram van te maken.

SPSS Het maken van grafieken

SPSS kan op verschillende manieren een grafiek voor je maken. De eenvoudigste manier is om dit te doen bij het maken van een frequentietabel. Via *Analyze → Descriptive Statistics → Frequencies* geef je eerst aan dat je een frequentieverdeling wilt maken. In het betreffende scherm (zie kader 1.1 figuur B) kun je nu Charts aanklikken. Vervolgens kun je in het Charts-venster (zie figuur A) aangeven of je een staafdiagram (Bar charts), een taartdiagram (Pie charts) of een histogram (Histograms) wilt. Histogrammen zullen besproken worden in hoofdstuk 3. Onder Chart Values geef je aan of dit in absolute frequenties of in percentages moet worden weergegeven.
Door in de output van SPSS dubbel te klikken op het figuur is het nog mogelijk om de tekst te bewerken of de arceringen te veranderen.



Figuur A Charts-venster

Een andere manier om grafieken te maken is via Graphs. Deze optie geeft veel verschillende soorten grafieken, die hier niet nader worden besproken.

Kader 1.2

1.3 Kruistabellen

Tot nu toe hebben we bij het maken van frequentieverdelingen steeds naar één enkele variabele afzonderlijk gekeken. Je kunt ook twee variabelen tegelijk gebruiken in je analyses. Wanneer je dat doet, maak je bijvoorbeeld een *kruistabel*. In feite ga je weer turven, maar nu gebruik je de waarden van twee variabelen tegelijkertijd. Hoeveel vrouwen zijn er 19 jaar? En hoeveel mannen? In tabel 1.6 is aan de hand van de datamatrix (tabel 1.5) een kruistabel gemaakt van de variabelen leeftijd en sekse. Van de twee 19-jarigen is er één man en één vrouw. De twee 20-jarigen zijn vrouwen. Zo ga je het hele rijtje af.

Tabel 1.6 Verdeling van leeftijd naar sekse (absolute frequenties)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	1	1	2
20	2	0	2
21	0	1	1
22	1	0	1
24	0	1	1
Totaal	4	3	7

Ook bij kruistabellen is het overzichtelijk om de percentages erbij te geven. Dat kan in een kruistabel op drie manieren.

Totaalpercentages
Je stelt het totaal aantal onderzoekseenheden op 100% en berekent dan de celpercentages.

Tabel 1.7 Verdeling van leeftijd en sekse, percentages (gepercenteerd op totaal)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	14,3	14,3	28,6
20	28,6	0	28,6
21	0	14,3	14,3
22	14,3	0	14,3
24	0	14,3	14,3
Totaal	57,1	42,9	100

Je kunt uit deze tabel aflezen dat 14,3% van alle onderzochte personen 19 jaar en vrouw is en 14,3% 19 jaar en man is. Op deze manier interpreteer je alle percentages in deze tabel.

Kolompercentages
De tweede manier om percentages in een kruistabel te berekenen is door de onderzoekseenheden in de kolommen op 100% te stellen en dan de percentages te berekenen.

Tabel 1.8 Verdeling van leeftijd naar sekse, percentages (gepercenteerd op sekse)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	25,0	33,3	28,6
20	50,0	0	28,6
21	0	33,3	14,3
22	25,0	0	14,3
24	0	33,3	14,3
Totaal	100	100	100

In tabel 1.8 is sekse, de kolomvariabele, op 100% gesteld. De percentages in de kolommen tellen op tot 100%. Je ziet dat deze kolompercentages in de cellen anders zijn dan de percentages in tabel 1.7, de interpretatie is ook anders. Je redeneert vanuit de onderzoekseenheden die op 100% zijn gesteld: 25% van alle vrouwen is 19 jaar en 33,3% van alle mannen is 19 jaar. In dit boek gebruiken we in de regel kolompercentages.

Rijpercentages
De laatste manier om percentages in een kruistabel te berekenen is door de onderzoekseenheden in de rijen op 100% te stellen, en dan de percentages te berekenen. Wanneer je leeftijd, de rijvariabele, tot 100% laat optellen, krijg je weer andere percentages in de kruistabel, met weer een andere interpretatie.

Tabel 1.9 Verdeling van sekse naar leeftijd, percentages (gepercenteerd op leeftijd)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	50	50	100
20	100	0	100
21	0	100	100
22	100	0	100
24	0	100	100
Totaal	57,1	42,9	100

Uit tabel 1.9 blijkt dat 50% van alle 19-jarigen vrouw is. Van alle 21-jarigen is 100% man.

Welke manier van percenteren je kiest, is afhankelijk van welke uitspraken je wilt doen op basis van je onderzoek. Als je iets wilt zeggen over de man-vrouw-verdeling naar leeftijd, dan moet je percenteren over de variabele leeftijd. Als je iets wilt zeggen over de leeftijden van mannen in vergelijking met vrouwen, percenteer je over de variabele sekse. Als er sprake is van een afhankelijke en een onafhankelijke variabele, percenteer je over de onafhankelijke variabele[1] (zie paragraaf 1.4).

Hoe ziet dit eruit in SPSS?
Je kunt SPSS kruistabellen laten maken met absolute frequenties en alle drie de genoemde percentuele berekeningen. Dan krijg je in elke cel erg veel getallen. Daarom is het beter om een keuze te maken en aan te geven op welke variabele je wilt percenteren (welke variabele je op 100% zet). De output in tabel 1.10 komt overeen met tabel 1.8 (gepercenteerd op sekse, kolompercentages).

In een SPSS-kruistabel worden achter *Count* de absolute frequenties gegeven. Er is één vrouw en één man van 19 jaar. In totaal zijn er twee 19-jarigen. In de onderste rij kun je lezen dat er in totaal vier vrouwen en drie mannen zijn. Onder *Count* staat '% within sekse'. Dit houdt in dat je gepercenteerd hebt naar sekse (ook te zien aan de 100% in de onderste rij van de kolommen). Van de vrouwen is 25% 19 jaar en van de mannen is 33,3% 19 jaar.

Als je in SPSS een kruistabel maakt, zie je de basiselementen terug zoals die besproken zijn in figuur 1.1. Je ziet bijvoorbeeld zowel de variabelenaam als het variabelelabel boven de kruistabel staan, en ook in de kolommen en de rijen komt deze informatie terug. In tabel 1.11 zie je bijvoorbeeld dat gekeken is of mannen en vrouwen van elkaar verschillen in de mate van interesse in kunst.

Tabel 1.10 Verdeling van leeftijd naar sekse (SPSS-output)

leeftijd * sekse Crosstabulation

			sekse		Total
			vrouw	man	
leeftijd	19	Count	1	1	2
		% within sekse	25,0%	33,3%	28,6%
	20	Count	2	0	2
		% within sekse	50,0%	,0%	28,6%
	21	Count	0	1	1
		% within sekse	,0%	33,3%	14,3%
	22	Count	1	0	1
		% within sekse	25,0%	,0%	14,3%
	24	Count	0	1	1
		% within sekse	,0%	33,3%	14,3%
Total		Count	4	3	7
		% within sekse	100,0%	100,0%	100,0%

Boven de kruistabel staat 'V57 interesse in kunst * V49 sekse Crosstabulation'. V57 is de variabelenaam van de variabele die als label 'interesse in kunst' heeft, geslacht heeft als variabelenaam 'V49' en als variabelelabel 'sekse'. De variabele die interesse in kunst meet heeft drie waarden, die waarden hebben als labels 'sterk in geïnteresseerd', 'tamelijk sterk in geïnteresseerd' en 'niet zo in geïnteresseerd'. De variabele die de sekse van de respondent meet, heeft twee waarden met als labels 'man' en 'vrouw'.

Tabel 1.11 Kruistabel van interesse in kunst naar sekse (SPSS-output)

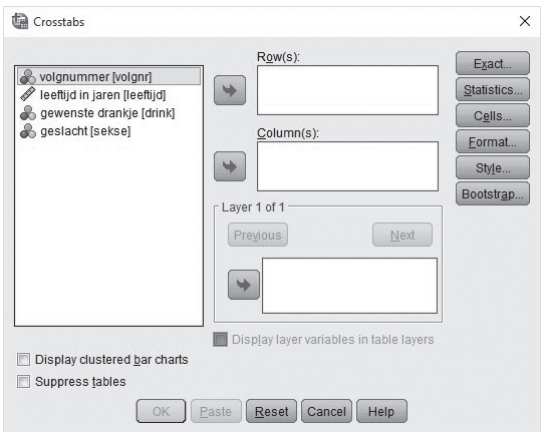
V57 Interesse in kunst *V49 sekse Crosstabulation

			V49 sekse		Total
			1 man	2 vrouw	
V57 Interesse in kunst	1 sterk in geïnteresseerd	Count	62	97	159
		& within V49 sekse	6,9%	10,3%	8,6%
	2 tamelijk sterk in geïnteresseerd	Count	190	267	457
		& within V49 sekse	21,2%	28,3%	24,9%
	3 niet zo in geïnteresseerd	Count	644	579	1223
		& within V49 sekse	71,9%	61,4%	66,5%
Total		Count	896	943	1839
		& within V49 sekse	100,0%	100,0%	100,0%

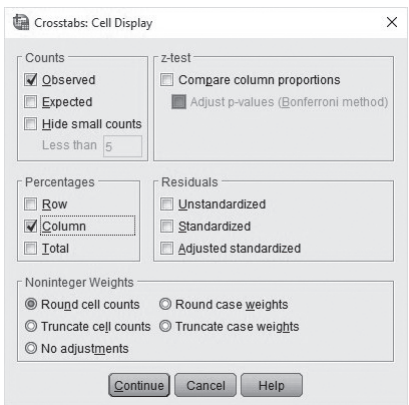
Bovenstaande kruistabel, die gepercenteerd is op de kolommen, laat bijvoorbeeld zien dat van de 896 mannen die aan deze enquête hebben deelgenomen 6,9% sterk in kunst is geïnteresseerd, tegenover 10,3% van de 943 vrouwen.

SPSS Kruistabellen maken

Om een kruistabel te maken in SPSS ga je via Analyze → Descriptive Statistics naar Crosstabs. In het Crosstabs-venster (zie figuur A) kun je aangeven welke variabele in de rijen (Row(s)) en welke variabele in de kolommen (Column(s)) moet komen. Doorgaans kiezen wij ervoor om de onafhankelijke variabele in de kolommen te zetten en de afhankelijke variabele in de rijen.
Om aan te geven of er op de kolommen of de rijen gepercenteerd moet worden, klik je op Cells, om vervolgens in de Cell Display (zie figuur B) aan te geven of je op de rijen of de kolommen wilt percenteren. Via Format kun je eventueel aangeven dat je de waarden op- of aflopend wilt presenteren.



Figuur A Crosstabs-venster



Figuur B Cell Display-venster