Beschrijvende statistiek

Het berekenen en interpreteren van tabellen en statistieken

Bregje van Groningen
Connie de Boer

Vierde druk

Boom

inclusief website!

Met behulp van onderstaande unieke activeringscode kunt u toegang krijgen tot www.beschrijvendestatistiek.nl voor extra materiaal. Deze code is persoonsgebonden en gekoppeld aan de 4e druk. Na activering van de code is de website 2 jaar toegankelijk. De code kan tot zes maanden na het verschijnen van een volgende druk geactiveerd worden.

9823-XL-53-LG

Omslagontwerp: Cunera Joosten, Amsterdam
Opmaak binnenwerk: Nu-nique grafische vormgeving, Goor

ISBN 9789089539823
ISBN 9789461278432 (e-book)
NUR 916

www.boomuitgeversamsterdam.nl
www.beschrijvendestatistiek.nl

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.