STATISTIEK VOOR ECONOMIE

TOUW

Noordhoff Uitgevers

Statistiek voor economie

Statistiek voor economie

ir. P. Touw

Tweede druk

Noordhoff Uitgevers Groningen

Omslagontwerp:
Designstudio Rob Buschman BNO, Gees

Basisvormgeving binnenwerk:
Carla Gerritzen BNO, Breda

Omslagfoto:
Hollandse Hoogte, Amsterdam, Roel Santvoort

Eventuele op- en aanmerkingen over deze of andere uitgaven kunt u richten aan: Noordhoff Uitgevers bv, Afdeling Hoger Onderwijs, Antwoordnummer 13, 9700 VB Groningen, e-mail: info@noordhoff.nl

3 4 5 / 13 12 11 10 09

ISBN (ebook) 978-90-01-84904-7
ISBN 978-90-01-87148-2
NUR 815

Woord vooraf bij de tweede druk

De tweede druk van de methode *Statistiek voor economie* borduurt voort op de eerste druk. Naast de modernisering van de lay-out is de tekst daar waar nodig geactualiseerd. De belangrijkste inhoudelijke veranderingen zijn de introductie van de software-pakketten Excel en SPSS en de toevoeging van een apart toepassingenboek.

De methode *Statistiek voor economie* heeft het volgende te bieden:

- Elk hoofdstuk begint met een probleemomschrijving (een case) vanwaaruit een probleemstelling wordt geformuleerd. Om tot een oplossing te komen, worden statistische technieken besproken die daarvoor kunnen worden aangewend.
- De nadruk ligt op toegepaste statistiek, waarbij door het veelvuldig gebruik van krantenartikelen en herkenbare praktijkvoorbeelden steeds de koppeling tussen theorie en praktijk wordt gelegd. Denk daarbij bijvoorbeeld aan toepassingsgebieden binnen marktonderzoek en accountantscontrole.
- Het boek geeft de student de gelegenheid zelfstandig de leerstof te bestuderen. Dit wordt bereikt doordat in de tekst een terugkoppeling aanwezig is in de vorm van korte vragen, waardoor de student steeds aan het denken wordt gezet. De lestijd kan zodoende efficiënter worden ingevuld met individuele begeleiding en het bespreken van de opgaven.
- Bij de vragen en de opgaven wordt – indien van toepassing – verwezen naar software-pakketten die als hulpmiddel kunnen dienen bij de berekeningen. Er wordt gebruikgemaakt van het spreadsheet-pakket Excel en van het statistische pakket SPSS.
- In een apart toepassingenboek zijn per hoofdstuk opgaven opgenomen, met de bijbehorende antwoorden. Deze opgaven bieden de student extra oefenstof. Bij het toepassingenboek is een diskette gevoegd, met Excel- en SPSS-bestanden die de uitwerking van vragen en opgaven illustreren.
- Naast het theorieboek en het toepassingenboek is een docentenhandleiding op cd-rom beschikbaar. Hierin zijn de uitwerkingen van de opgaven uit het toepassingenboek (inclusief Excel- en SPSS-bestanden) opgenomen.

Ik hoop dat de genoemde wijzigingen ertoe leiden dat er nog succesvoller met het materiaal kan worden gewerkt. Reacties op de methode zijn van harte welkom: info@epn.nl.

Ten slotte wil ik mijn naaste collega, drs. R.L. Erven, bedanken voor zijn suggesties en zijn opbouwende kritiek.

Zwolle, voorjaar 2000
ir. P. Touw

Studiewijzer

Indeling van het boek

Het boek is opgebouwd uit zes hoofdstukken van in totaal 80 studiebelastingsuren (80 sbu). Hieronder volgt een korte beschrijving van de inhoud van ieder hoofdstuk.

In *hoofdstuk 1 Gegevens presenteren* (± 15 sbu) wordt aan de hand van een enquête over een bibliotheek gekeken hoe onderzoeksresultaten getalsmatig en grafisch kunnen worden gepresenteerd. Hierbij wordt gebruikgemaakt van de statistische mogelijkheden van de rekenmachine.

In *hoofdstuk 2 Kansrekening* (± 15 sbu) wordt het begrip kans en de bijbehorende kansregels uitgelegd aan de hand van de uitwerking van verschillende kleine voorbeelden. Tevens wordt het begrip kansvariabele ingeleid, dat dient als voorbereiding op het volgende hoofdstuk.

In *hoofdstuk 3 De binomiale verdeling* (± 10 sbu) wordt uitgelegd in welke gevallen en op welke manier kansen kunnen worden berekend met behulp van de binomiale verdeling en de cumulatieve binomiale tabel.

In *hoofdstuk 4 Lineaire regressie* (± 10 sbu) wordt op basis van de beschikbare gegevens met behulp van de kleinstekwadratenmethode de prijs-vraagfunctie berekend van de NS-voordeel-urenkaart. Ook wordt onderzocht hoe goed deze functie het verband tussen de prijs en de vraag beschrijft. Het hoofdstuk wordt afgesloten met een casestudy waarin het gasverbruik in Nederland wordt gerelateerd aan de temperatuur tijdens een koude periode.

In *hoofdstuk 5 Het seizoensmodel* (± 10 sbu) wordt op basis van de kwartaalcijfers van de KLM van enkele jaren een voorspelling gegeven van de nieuwe kwartaalcijfers. Hiertoe worden met behulp van het voortschrijdend gemiddelde de trend en met behulp van het seizoensmodel de seizoenscomponenten bepaald. Het hoofdstuk wordt afgesloten met een casestudy over de werkloosheid in Nederland.

In *hoofdstuk 6 De normale verdeling* (± 20 sbu) wordt gekeken naar de toepassingsmogelijkheden van de normale verdeling en op welke manier kansen met behulp van de tabel van de standaardnormale verdeling kunnen worden berekend. Tevens wordt de geldigheid van de normale verdeling toegelicht aan de hand van de centrale limietstelling en wordt daarvan een voorbeeld gegeven in de benadering van de binomiale verdeling door de normale verdeling. Ten slotte wordt gekeken naar de sommatie en het gemiddelde van normaal verdeelde variabelen.

Methodiek

Om de stof goed te kunnen volgen, dient de student de hoofdstukken in de gepresenteerde volgorde te doorlopen. Er is geen uitgebreide wiskundige voorkennis vereist. Afhankelijk van de vooropleiding (vwo, havo of mbo) en het vakkenpakket kan een student (in overleg met de docent) één of meer hoofdstukken overslaan. Deze hoofdstukken kunnen eventueel als naslagwerk worden geraadpleegd als een student in de rest van het boek problemen ondervindt.

Bij het schrijven van dit materiaal is voor elk hoofdstuk zoveel mogelijk de volgende methodiek toegepast. Een hoofdstuk begint met een opsomming van de leerdoelen van dat hoofdstuk. Daarna volgt een korte inleiding (bijvoorbeeld een krantenartikel), waardoor de lezer een indruk krijgt van de inhoud van het hoofdstuk.
Vanuit een praktische probleemomschrijving (case) wordt vervolgens een probleemstelling geformuleerd. In sommige gevallen is ervoor gekozen om door het geven van een grote hoeveelheid voorbeelden aan te geven, waar de toepassingsgebieden van de leerstof kunnen worden gevonden. Aan de hand van de probleemstelling worden statistische methoden aangedragen die tot een oplossing kunnen leiden.
Tijdens het bestuderen van de theorie wordt de student door het beantwoorden van korte vragen verder geholpen zich de stof eigen te maken. Aan het eind van het boek zijn de uitwerkingen van deze vragen opgenomen.
In de meeste gevallen is een aanvullende of een verdiepende case opgenomen. Bij de introductie van belangrijke begrippen worden deze begrippen in de marge vermeld. Deze kernbegrippen en de belangrijkste formules worden aan het eind van het hoofdstuk nog eens op een rijtje gezet.

Aan het eind van ieder hoofdstuk wordt de student verwezen naar het toepassingenboek. Door het maken van de opgaven in het toepassingenboek kan de student met de leerstof oefenen en nagaan of hij de theorie voldoende beheerst.

Bij de vragen in het theorieboek en de opgaven in het toepassingenboek is door middel van de symbolen (E) en (S) aangegeven of bij de beantwoording gebruik kan worden gemaakt van de software-pakketten Excel en SPSS.
In de uitwerkingen staat beschreven hoe dit gedaan kan worden.

Opzet van het boek

Bij het bestuderen van het boek kunnen natuurlijk gewoon hoofdstuk 1 tot en met hoofdstuk 6 achter elkaar worden bestudeerd. Is er echter behoefte om slechts een deel van de kennis op te halen, dan moet duidelijk zijn welke hoofdstukken daarvoor eerst bestudeerd moeten worden. In onderstaande tabel staat voor elk hoofdstuk aangegeven welke voorkennis voor dat hoofdstuk vereist is. Hoofdstuk 1 behoeft (nagenoeg) geen voorkennis. Hoofdstuk 5 is niet als voorkennis binnen een ander hoofdstuk vereist.

	hoofdstuk 1	*hoofdstuk 2*	*hoofdstuk 3*	*hoofdstuk 4*
hoofdstuk 2	gemiddelde staafdiagram standaardafwijking standaarddeviatie			
hoofdstuk 3	gemiddelde staafdiagram standaardafwijking standaarddeviatie	complementregel onafhankelijk productregel bijzondere somregel combinatie binomiaalcoëfficiënt kansfunctie kansvariabele stochast verwachtingswaarde		
hoofdstuk 4	gemiddelde standaardafwijking standaarddeviatie populatie steekproef	verwachtingswaarde		
hoofdstuk 5	gemiddelde			regressielijn kleinstekwadraten-methode correlatiecoëfficiënt
hoofdstuk 6	gemiddelde staafdiagram standaardafwijking standaarddeviatie	complementregel kansvariabele stochast verwachtingswaarde	binomiale verdeling binomiale formule binomiaalcoëfficiënt	

Hieruit volgt een noodzakelijke volgorde bij bestudering van:

- hoofdstuk 6: H1 → H2 → H3 → H6 (± 60 sbu)
- hoofdstuk 5: H4 → H5 (± 20 sbu)
- hoofdstuk 4: H1 → H4 (± 25 sbu)
- hoofdstuk 3: H1 → H2 → H3 (± 40 sbu)
- hoofdstuk 2: H1 → H2 (± 30 sbu)

Inleiding

Een korte historie

Het verzamelen van statistische gegevens is al minstens zo oud als de weg naar Rome, in werkelijkheid zelfs al veel ouder. Er is een kleitablet opgegraven uit de Oud-Babylonische periode (1600 tot 1900 voor Christus), waarop een statistische berekening is weergegeven. In de bijbel (Lucas 2) wordt reeds melding gemaakt van een volkstelling ten tijde van de geboorte van Jezus, op bevel van keizer Augustinus (63 voor Christus tot 14 na Christus).
De kansrekening is de eerste aanzet tot een meer theoretische onderbouwing van de huidige statistiek. Deze ontwikkeling kwam pas in de zeventiende eeuw op gang, in eerste instantie vooral op het gebied van de kansspellen. Bekende namen zijn:

Blaise Pascal (1623–1662)
Frans filosoof, wiskundige, natuurkundige en schrijver
Pierre de Fermat (1601–1665)
Frans wiskundige en jurist

Ook een bekende Nederlander heeft zich beziggehouden met de kansrekening. Zijn interesse lag echter meer op het gebied van de verzekeringswiskunde, waarbij premies voor lijfrentes werden berekend op basis van sterftekansen:

Christiaan Huygens (1620–1659)
Nederlands wiskundige, natuurkundige en astronoom

Een Zwitser hield zich bezig met de kansrekening in relatie tot meetfouten:

Jakob Bernoulli (1654–1705)
Zwitsers wiskundige

Een Franse geleerde kwam met een belangrijke ontdekking: de wiskundige formule van de normale verdeling:

Abraham de Moivre (1667–1754)
Frans wiskundige

Een belangrijke plaats in deze beginperiode van de statistiek is weggelegd voor een andere Frans geleerde, die de basis legde voor de formulering van de klassieke kansrekening:

Pierre-Simon Laplace (1749–1827)
Frans wiskundige en astronoom

Uit de negentiende en twintigste eeuw is nog een aantal namen te noemen met hun bekendste ontdekkingen:

Johann Carl Friedrich Gauβ (1777–1855)
Duitser, bekend geworden vanwege de praktische toepassing van de normale verdeling

Siméon Denis Poisson (1781–1840)
Fransman, bedenker van wat later de Poisson-verdeling is gaan heten

Karl Pearson (1857–1936)
Engelsman, bekend vanwege zijn correlatiecoëfficiënt

Ronald Aylmer Fisher (1890–1962)
Engelsman, heeft een bijdrage geleverd aan de variantieanalyse

Andrey Nikolayevich Kolmogorov (1903–1987)
Rus, bekend door zijn kansaxioma's

Statistische centra

Nederland beschikt over verschillende statistische centra. Het *Centraal Bureau voor de Statistiek* (*CBS*) geniet de meeste bekendheid vanwege de uitgave van het *Statistisch Jaarboek*.
De *Vereniging voor Statistiek* (*VVS*) organiseert congressen op het gebied van de toegepaste statistiek. Op veel (technische) universiteiten wordt ook onderzoek gedaan naar statistiek.
Internet biedt een ongekend aantal mogelijkheden om statistische gegevens te 'downloaden'. Ook vind je er informatie over cursussen en recente ontwikkelingen binnen de statistiek. Een kleine greep uit de beschikbare websites:

De Digitale School
http://www.digischool.nl/wi/wikrstat.htm

Centraal Bureau voor de Statistiek
http://www.cbs.nl

Eurostat
http://europa.eu.int/comm/eurostat

U.S. Department of Commerce
http://www.esa.doc.gov

Southwest Missouri State University
http://www.psychstat.smsu.edu/script/dws148f/statisticsresources-main.asp
(een leuke 'site' met links naar o.a. On-line Statistics Books, Articles, Data Files)

Inhoud

Hoofdstuk
1

Gegevens presenteren

‘There are three kinds of lies: lies, damned lies and statistics.’

Benjamin Disraeli (1804–1881)

‘There are two kinds of statistics, the kind you look up, and the kind you make up.’

Rex Stout (1886–1975)

LEERDOELEN

Na bestudering van dit hoofdstuk is de student in staat om:

- ***aan te geven dat de gegevens van een nominale schaal, een ordinale schaal of een ratioschaal zijn;***
- ***de modus, de mediaan en het gemiddelde te berekenen bij ‘losse’ waarnemingen en bij een klasse-indeling;***
- ***de spreidingsbreedte en de standaardafwijking te berekenen bij ‘losse’ waarnemingen en bij een klasse-indeling;***
- ***de variatiecoëfficiënt te berekenen;***
- ***afhankelijk van de aard van de gegevens en de probleemstelling een keuze te maken uit (een of meer van) de volgende grafieken: staafdiagram, stapeldiagram, histogram, frequentiepolygoon, (relatief) cumulatief frequentiepolygoon en cirkeldiagram;***
- ***uit een grafiek verbanden af te lezen.***

Inleiding

In deze tijd van informatietechnologie zijn er talloze mogelijkheden om informatie in te winnen. Zo hebben bijvoorbeeld het Centraal Bureau voor de Statistiek (CBS), de Kamer van Koophandel en vele bedrijfstakken eigen informatiediensten. Soms is de gewenste informatie echter niet beschikbaar, te oud of niet (geheel) van toepassing op de doelgroep. Dan kan men overwegen zelf een onderzoek uit te voeren.
De verwerking van de gegevens gebeurt tegenwoordig bijna alleen nog maar met statistische software-pakketten. Het gebruik van een computer kan veel werk besparen. Om de gegevens op de juiste manier te verwerken en te presenteren moet inzicht worden verkregen in de aard van de gegevens en kennis worden opgedaan van de beschikbare statistische presentatietechnieken.
In dit hoofdstuk zullen aan de hand van een enquête verschillende statistische begrippen de revue passeren.

1.1 Case: De Bibliotheek Barometer

Om staande te blijven in een tijd met sterk veranderende media heeft de 'ouderwetse' bibliotheek met louter boeken al een behoorlijke gedaanteverwisseling ondergaan. Nog steeds bestaat de kernactiviteit uit het uitlenen van boeken, maar er is duidelijk sprake van een accentverschuiving. Er is ruimte gemaakt voor een fonotheek, waar cd's (en misschien nog enkele lp's) staan uitgestald, er zijn video's van bekende tekenfilms en kinderfilms, er zijn video's van klassiekers voor volwassenen en er is een experiment gaande met cd-roms.

VOORBEELD

De Bibliotheek Barometer van het NBLC (Nederlands Bibliotheek- en Lectuurcentrum) is opgezet vanuit de gedachte dat kennis over het oordeel van de gebruikers over de dienstverlening van de bibliotheek een noodzakelijke voorwaarde is voor een effectief bibliotheekbeleid. Gebruikersonderzoek is een van de instrumenten daartoe en wordt bij bibliotheken al jaren toegepast ...
... De 'Bibliotheek Barometer' is een standaardgebruikersonderzoek waarmee vergelijkingen kunnen worden gemaakt met de resultaten van andere bibliotheken. Door het standaardkarakter van het onderzoek en doordat de bibliotheek zelf het 'veldwerk' verzorgt kunnen de kosten van het onderzoek worden gedrukt.

Voor het onderzoek is een standaardvragenlijst ontwikkeld waarmee met name de oordelen van de klant over bijna alle aspecten van de dienstverlening een prominente plaats innemen. Daarnaast is er nog een aantal vragen opgenomen over het bibliotheekgebruik en worden er natuurlijk vragen naar de meest gebruikelijke sociaal-economische kenmerken van de gebruikers gesteld. Doordat het mogelijk is extra vragen op te nemen, kan de Bibliotheek Barometer aangepast worden aan de eigen lokale situatie ...
... De resultaten van het Barometeronderzoek kunnen gebruikt worden als ondersteuning van het gevoerde en voorgenomen bibliotheekbeleid. Het onderzoek leent zich namelijk ook uitstekend voor een vergelijking in de tijd; met een herhaling van het onderzoek kunnen de effecten van veranderingen in de dienstverlening op de tevredenheid van de klanten gemeten worden.

Bibliotheek Zwolle, voorjaar 1996

Hieronder is een aantal vragen uit een vergelijkbare enquête opgenomen.

1 Van welk geslacht bent u?

☐ man
☐ vrouw

2 Hoe oud bent u?

☐ minstens 10 en jonger dan 15
☐ minstens 15 en jonger dan 25
☐ minstens 25 en jonger dan 35
☐ minstens 35 en jonger dan 50
☐ minstens 50 en jonger dan 65
☐ minstens 65

3 Van welke voorziening van de bibliotheek maakt u geregeld gebruik?

	nooit	1	2	3	4	5	altijd
• informatiebalie		☐	☐	☐	☐	☐	
• computerinformatiesysteem		☐	☐	☐	☐	☐	
• sanitaire voorzieningen		☐	☐	☐	☐	☐	
• fietsenstalling		☐	☐	☐	☐	☐	

4 Van welke afdeling maakt u geregeld gebruik?

	ja	nee
• studiezaal	☐	☐
• volwassenenafdeling	☐	☐
• leeszaal	☐	☐
• fonotheek	☐	☐
• jeugdafdeling	☐	☐

5 Hoeveel geld besteedt u maandelijks aan de koop van boeken, tijdschriften en kranten?

.....

6 Wat is uw algemene oordeel over de bibliotheek in zijn geheel?
- ☐ zeer tevreden
- ☐ tevreden
- ☐ neutraal
- ☐ ontevreden
- ☐ zeer ontevreden

Een belangrijk statistisch probleem dat buiten het bestek van dit boek valt is de vraag onder hoeveel klanten van de bibliotheek de enquête moet worden verspreid (tien personen, vijftig personen of meer?).
Een ander probleem is hoe de enquête moet worden uitgevoerd. Een mogelijkheid is om de formulieren te versturen naar adressen van mensen die lid zijn. Ook kan men de enquête telefonisch afnemen of kan men overwegen om de klanten in de bibliotheek te ondervragen. Deze vragen komen aan bod bij het vakgebied marktonderzoek.
Het marktonderzoeksbureau dat deze opdracht uitvoert besluit (in overleg met de directie van de bibliotheek) tot een mondelinge enquête onder 150 klanten bij de uitgang van de bibliotheek. Bij een mondelinge enquête is het mogelijk iemand persoonlijk te helpen met de invulling van de vragenlijst. Tevens verkrijgt men een goed beeld van alle klanten van de bibliotheek. Dit zijn niet alleen leden, maar ook niet-leden die toch gebruikmaken van de voorzieningen.

1.1.1 Losse waarnemingen

Om te onderzoeken of de enquête goed in elkaar steekt wordt deze eerst onder tien bezoekers van de bibliotheek getest. Het resultaat van die proef is weergegeven in figuur 1.1. Naast de vragen staat een korte omschrijving. Op elke vraag kunnen verschillende antwoorden worden gegeven.

variabele

Men noemt een vraag ook wel een variabele die diverse waarden kan aannemen. Vraag 1 wordt bijvoorbeeld de variabele ‘geslacht’ en kan de waarden m (man) en v (vrouw) aannemen. Vraag 3 en vraag 4 zijn op te splitsen in respectievelijk vier en vijf subvragen. Op deze manier ontstaan in totaal dertien variabelen.
Het is in de praktijk gebruikelijk om alle uitkomsten als getallen te coderen. Voor vraag 1 ‘geslacht’ zou dit bijvoorbeeld kunnen betekenen dat voor ‘man’ een ‘1’ wordt ingevuld en voor ‘vrouw’ een ‘2’. Statistische software kan met ingevoerde getallen meer analyses uitvoeren dan met letters.

Voor de leesbaarheid van de resultatenmatrix is gekozen om deze coderingen achterwege te laten.

vraag	*omschrijving variabele*	*respondent* *1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	geslacht	v	m	m	m	m	v	v	v	v	v
2	leeftijd	15/25	≥ 65	15/25	35/50	15/25	15/25	35/50	50/65	50/65	15/25
3	infobalie	1	4	2	1	1	2	2	4	3	3
	computerinfo	4	1	5	2	4	3	4	1	1	4
	sanitair	1	2	3	1	2	2	5	5	4	1
	fietsenstalling	5	1	5	1	2	2	2	5	4	5
4	studiezaal	j	n	n	j	n	n	j	n	j	n
	volwassenen	j	j	n	j	n	j	j	j	j	j
	leeszaal	n	j	n	n	n	n	j	n	j	j
	fonotheek	j	n	j	j	j	j	j	j	n	n
	jeugdafdeling	j	n	j	j	j	n	j	n	n	j
5	geld	27,5	2,5	12,5	45	30	12,5	22,5	15	40	17,5
6	oordeel	zt	zt	o	zt	n	zt	n	t	t	n

Figuur 1.1 *Resultatenmatrix van antwoorden van tien personen*

PROBLEEMSTELLING Op welke manieren kunnen de voorlopige resultaten worden gepresenteerd?

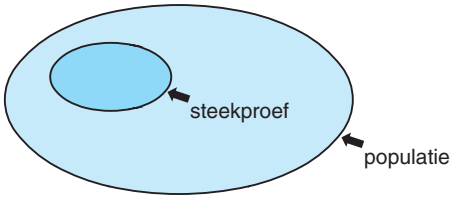
De mondelinge enquête die is afgenomen onder de tien bezoekers van de bibliotheek geeft een eerste indruk van de mening van de bezoekers en hoe de vragen in de praktijk worden opgevat. In de groep zijn vier mannen aangetroffen en zes vrouwen.

VRAAG 1.1 Is het gerechtvaardigd om te concluderen dat 40 procent van de klanten van de bibliotheek man is?

steekproef

populatie

De groep van tien klanten wordt een steekproef genoemd. Een steekproef is een willekeurige selectie van klanten uit het totaal van klanten. Het totaal van klanten van de bibliotheek noemt men de populatie. In figuur 1.2 is het verschil tussen een steekproef en een populatie schematisch weergegeven.



Figuur 1.2
Steekproef en populatie

Er zijn vele methoden ontwikkeld voor het nemen van steekproeven, die afhangen van de complexiteit van de populatie en de beschikbare financiën (en tijd). Al deze methoden zorgen ervoor dat de steekproef een goede afspiegeling is van de populatie. Als dat het geval is noemt men de steekproef 'representatief'.
Als we willen dat elke klant van de bibliotheek dezelfde kans heeft om voor de steekproef te worden geselecteerd dan noemt men de steekproef aselect.

aselect

1.1.1.1 Maten van ligging (centrummaten)

De vragen die in de mondelinge enquête zijn gesteld kunnen worden ondergebracht in verschillende soorten schalen. In de laagste schaal zijn er geen getalwaarden en is er geen volgorde aan te brengen tussen de mogelijke antwoorden. Een voorbeeld hiervan is vraag 1: 'Van welk geslacht bent u?' De antwoorden op deze vraag behoren tot de nominale schaal. Dus de bijbehorende variabele 'geslacht' is een nominale variabele.

nominale schaal

Wil je iets zeggen over de waarde van de variabele 'geslacht', dan is de enige mogelijkheid te vermelden welke antwoordmogelijkheid het meeste voorkomt. Het antwoord dat het meeste voorkomt wordt modus genoemd. Dit is een centrummaat.

modus

NOTATIE Mo

Voor de variabele 'geslacht' is: Mo = vrouw.

VRAAG 1.2 Is er in de enquête nog een variabele te vinden die behoort tot de nominale schaal?

In de volgende schaal is er nog geen sprake van getalwaarden, maar er kan wel een volgorde worden aangebracht in de antwoordmogelijkheden. Een voorbeeld hiervan is vraag 6: 'Wat is uw algemene oordeel over de bibliotheek in zijn geheel?' De variabele 'oordeel' kan de waarden aannemen die variëren tussen zeer tevreden en zeer ontevreden. De variabele 'oordeel' behoort tot de ordinale schaal.

ordinale schaal

mediaan

Naast de modus kan van een ordinale variabele als centrummaat de mediaan worden berekend. De mediaan is de middelste van de waarneming en op volgorde van grootte.

NOTATIE Me

S VRAAG 1.3 Bepaal de modus van de variabele 'oordeel'.

Voor het bepalen van de mediaan moeten de antwoorden op volgorde van grootte worden gezet (van hoog naar laag of andersom, dit maakt niet uit).

zt, zt, zt, zt, **t, t,** n, n, n, o

Voor een oneven groep is het makkelijk de middelste waarneming aan te wijzen. Een even groep, als hierboven, heeft twee middelste waarnemingen, namelijk 'tevreden' en 'tevreden'. De mediaan is dan: Me = tevreden.

Opmerking
Bij een even groep is het mogelijk dat de middelste twee waarnemingen in verschillende categorieën vallen. Bijvoorbeeld:

zt, t, **t, n,** n, o

In dat geval kan de mediaan zijn: Me = neutraal/tevreden. Bij grotere steekproeven doet zich dit probleem nauwelijks voor.

VRAAG 1.4 Is er in de enquête nog een variabele met een ordinale schaal aanwezig?

ratioschaal

De hoogste schaal bestaat uit getalwaarden. Dit is de ratioschaal. Een voorbeeld hiervan is vraag 5: 'Hoeveel geld besteedt u maandelijks aan de koop van boeken, tijdschriften en kranten?' De bijbehorende variabele 'geld' bestaat uit tien getallen.

E S VRAAG 1.5 Bepaal de modus en de mediaan van de variabele 'geld'.

gemiddelde

Naast de modus en de mediaan kan het gemiddelde worden berekend.

NOTATIE $\bar{x}$

Het gemiddelde kan worden berekend door alle waarnemingen op te tellen en te delen door het aantal.

$$\bar{x} = \frac{27{,}5 + 2{,}5 + 12{,}5 + 45 + 30 + 12{,}5 + 22{,}5 + 15 + 40 + 17{,}5}{10}$$

$$= \frac{225}{10} = 22{,}5 \text{ euro per maand}$$

FORMULE 1.1

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{\sum x_i}{n}$$

waarbij:
x_i: de i-de waarneming
n: het aantal waarnemingen
$\sum$: de sommatie van

VRAAG 1.6 Is er nog een andere vraag met een ratioschaal in de enquête aanwezig?

Het nadeel van de formulering van de antwoorden van vraag 2 uit de enquête is dat de antwoorden alleen in leeftijdsklassen worden verkregen, zodat de bepaling van het gemiddelde een lastige klus wordt.

Opmerking
In de literatuur wordt naast de eerder genoemde schalen (nominaal, ordinaal en ratio) nog melding gemaakt van de intervalschaal. Evenals de ratioschaal bestaat deze uit getalwaarden. Het verschil met de ratioschaal is dat bij de intervalschaal waarnemingen alleen kunnen worden vergeleken aan de hand van verschillen en niet aan de hand van verhoudingen. Een voorbeeld van de intervalschaal is de temperatuur. Verschillen in temperatuur kunnen precies worden vastgesteld, maar het is niet zo dat 20 °C tweemaal zo warm is als 10 °C. De intervalschaal komt in de praktijk weinig of niet voor. Bij een intervalschaal kunnen de modus, de mediaan en het gemiddelde worden bepaald.

Figuur 1.3 geeft een overzicht van de schalen en de centrummaten die kunnen worden berekend.

	centrummaat		
schaal	*modus*	*mediaan*	*gemiddelde*
nominaal	ja	nee	nee
ordinaal	ja	ja	nee
ratio	ja	ja	ja

Figuur 1.3 *Centrummaten en schalen*

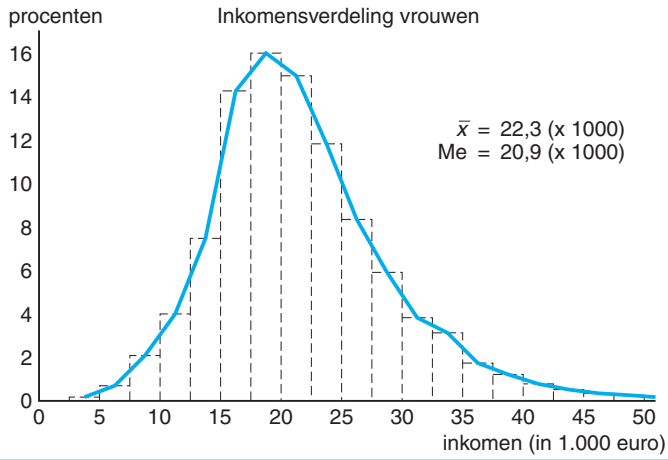
Doordat bij de ratioschaal de keuze bestaat om de variabele te beschrijven met drie centrummaten, gaat de voorkeur naar de ratioschaal uit. In het

algemeen wordt het gemiddelde als centrummaat het meest gebruikt. Het voordeel van het gemiddelde is dat elke waarneming even zwaar meetelt. In sommige gevallen ligt het gemiddelde niet voor de hand, bijvoorbeeld als er gegevens zijn met uitschieters.

VOORBEELD 1.1

Inkomensverdeling

Uit het *Statistisch Jaarboek 1996* (CBS) komt de volgende inkomensverdeling van vrouwen met een voltijdbaan (figuur 1.4).



Figuur 1.4 *Een asymmetrische inkomensverdeling* Bron: gebaseerd op Leeflang, 1986

In het geval van een scheve verdeling (uitschieters) wordt vaak gewerkt met modus en mediaan. Voor een inkomensverdeling is het bijvoorbeeld gebruikelijk te werken met het modale inkomen (Jan Modaal).

VRAAG 1.7

Geef een omschrijving van het modale inkomen (zie voorbeeld 1.1). Zal dit hoger of lager zijn dan het gemiddelde inkomen?

1.1.1.2 Maten van spreiding (spreidingsmaten)

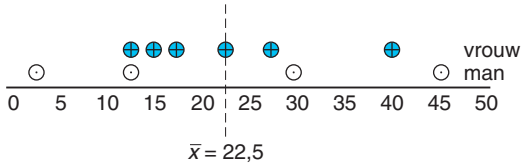
Bij een nominale variabele en een ordinale variabele is het niet mogelijk te berekenen hoeveel spreiding in de steekproef aanwezig is. Een ratioschaal biedt voldoende gegevens om te bepalen hoe groot de spreiding in de waarnemingen is.

Voor mannen en voor vrouwen (uit de steekproef van 10 personen) kan voor de variabele 'geld' worden berekend hoeveel men per groep per persoon gemiddeld besteedt (zie figuur 1.5).

groep	*man*	*vrouw*
	2,5	27,5
	12,5	12,5
	45	22,5
	30	15
		40
		17,5
totaal (€ per maand)	90	135
aantal	4	6
gemiddelde (€ per maand)	22,5	22,5

Figuur 1.5 *Gemiddelde besteding voor de groepen man en vrouw*

Beide groepen besteden in deze steekproef gemiddeld evenveel. Als we de waarnemingen uitzetten op een getallenlijn zien we het beeld van figuur 1.6.



Figuur 1.6 *Gemiddeld € 22,5 per maand*

VRAAG 1.8 Wat is het kenmerkende verschil tussen de groepen man en vrouw als wordt gekeken naar de variabele 'geld'?

spreidingsbreedte
range

Blijkbaar is het gemiddelde alleen als maat niet genoeg om een groep van waarnemingen te beschrijven. De meest eenvoudige spreidingsmaat is de spreidingsbreedte ofwel range. Dit is het verschil tussen de hoogste en de laagste waarneming.

NOTATIE *R*

Voor de maandelijkse besteding door mannen en vrouwen is dit:

Mannen : R = 45 − 2,5 = € 42,5 per maand
Vrouwen: R = 40 − 12,5 = € 27,5 per maand

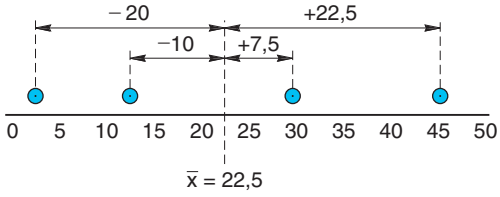
VRAAG 1.9 Wat zou het nadeel kunnen zijn van het gebruik van de spreidingsbreedte als maat van spreiding?

standaardafwijking
standaarddeviatie

De meest gebruikte spreidingsmaat is de standaardafwijking ofwel standaarddeviatie. Spreiding wordt veroorzaakt doordat een waarneming een bepaalde afstand van het gemiddelde af ligt.

VOORBEELD 1.2

Mannen en maandelijkse besteding



Figuur 1.7 *Afstand tot het gemiddelde*

Een waarneming onder het gemiddelde geeft evenveel spreiding als een waarneming boven het gemiddelde. Alle verschillen zijn gemakkelijk positief te maken door ze te kwadrateren.

Figuur 1.8 *Kwadratische afstand tot het gemiddelde*

waarneming (x_i)	*afstand tot gemiddelde* $x_i - \bar{x}$	*kwadratische afstand* $(x_i - \bar{x})^2$
2,5	−20	400
12,5	−10	100
30	7,5	56,25
45	22,5	506,25
totaal		1 062,5

In totaal geeft dit voor alle waarnemingen samen: 1 062,5.

Een sommatie kan worden weergegeven met een sommatieteken Σ.
$\Sigma (x_i - \bar{x})^2 = 1\,062{,}5$
Als er meer waarnemingen komen wordt deze sommatie alleen maar groter. Het is noodzakelijk om te 'middelen'.

Voor een steekproef moet dit nog worden gedeeld door het aantal min één.

$$\frac{1\,062{,}5}{4-1} = \frac{1\,062{,}5}{3} = 354{,}17$$

VRAAG 1.10 Wat is de eenheid van 354,17?

variantie

De waarde 354,17 noemt men in de statistiek de variantie.

NOTATIE s^2

FORMULE 1.2
$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Het nadeel van de variantie is dat je je bij de eenheid ervan weinig voor kunt stellen (zie de uitwerking bij vraag 1.10).

Als van de variantie nog de wortel wordt genomen dan wordt de standaardafwijking verkregen.

FORMULE 1.3
$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Dat geeft dus:

$s = \sqrt{s^2} = \sqrt{354{,}17} = 18{,}82$ euro per maand

De eenheid van standaardafwijking is weer dezelfde als die van het gemiddelde.

(S) VRAAG 1.11 Bereken de standaardafwijking voor de variabele 'geld' voor de groep vrouw.

Uit vraag 1.11 volgt dat de spreiding binnen de groep vrouw (€ 10,12) iets meer dan de helft is van de spreiding in de groep man (€ 18,82).

Tegenwoordig zijn bovenstaande berekeningen vrij gemakkelijk uit te voeren op een rekenmachine. Het voert te ver om voor alle typen een instructie op te nemen. Daarom voor een veel gebruikt (en goedkoop) apparaat een voorbeeld.

VOORBEELD 1.3

Casio fx-82SX Fraction

MODE	.	(zet de machine op statistiek)
SHIFT	AC	(maak het statistisch geheugen leeg)
2,5	M+	(voer het eerste getal in)
12,5	M+	
30	M+	
45	M+	(voer het laatste getal in)
SHIFT	6	(= 4 dit is ter controle het aantal waarnemingen)
SHIFT	7	(= 22,5 dit is het steekproefgemiddelde)
SHIFT	9	(= 18,82 dit is de steekproefstandaardafwijking)
MODE	0	(zet de machine op rekenen)

VOORBEELD 1.4

Spreiding in inkomen

In twee landen (Ivoorkust en Frankrijk) wordt een (even grote) steekproef afgenomen om het gemiddelde inkomen te bepalen. Om de inkomens te vergelijken worden de bedragen omgerekend naar dollars. In Ivoorkust was het resultaat: een gemiddelde van $ 690 met een standaardafwijking van $ 315. In Frankrijk was het resultaat: een gemiddelde van $ 20.600 met een standaardafwijking van $ 315.

VRAAG 1.12 Kan nu worden geconcludeerd dat in beide landen de spreiding in inkomen even groot is?

Het is niet ongebruikelijk om de standaardafwijking te vergelijken met het gemiddelde en dit eventueel in een percentage om te rekenen. Dit noemt men de variatiecoëfficiënt.

variatiecoëfficiënt

NOTATIE V

FORMULE 1.4 $V = \dfrac{s}{\bar{x}}$

Voor Ivoorkust en Frankrijk uit voorbeeld 1.4 geeft dit de volgende resultaten.

$$\text{Ivoorkust: } V = \frac{315}{690} = 0{,}457 \rightarrow 45{,}7\ \%$$

$$\text{Frankrijk: } V = \frac{315}{20.600} = 0{,}015 \rightarrow 1{,}5\ \%$$

De relatieve spreiding van het inkomen is voor Ivoorkust veel groter dan voor Frankrijk.

VRAAG 1.13 Bereken de waarde van de variatiecoëfficiënt van de maandelijkse besteding door mannen en door vrouwen aan boeken, tijdschriften en kranten (zie figuur 1.5).

1.1.2 Frequentieverdeling

Na deze vingeroefeningen is het nu tijd om naar de volledige steekproef van 150 klanten te kijken.

PROBLEEMSTELLING Op welke manieren kunnen de resultaten van de volledige steekproef worden gepresenteerd?

Het voert te ver om alle 'losse' waarnemingen te noteren, zoals bijvoorbeeld in figuur 1.1 is gedaan. Het is gebruikelijk om de resultaten weer te

frequentieverdeling

geven door middel van een klasseindeling ofwel een frequentieverdeling. Voor vraag 1 van de enquête: 'Van welk geslacht bent u?' is het resultaat:

- man: 54
- vrouw: 96

S VRAAG 1.14 Bepaal (indien mogelijk) de centrummaten en de spreidingsmaten.

Voor vraag 2: 'Hoe oud bent u?' zijn de resultaten van de steekproef weergegeven in figuur 1.9. Bekend is dat de jongste respondent tien jaar oud was en de oudste was 84 jaar. Het aantal waarnemingen per klasse

frequentie

noemt men de frequentie.

leeftijd	*frequentie*
10 – < 15	10
15 – < 25	32
25 – < 35	29
35 – < 50	40
50 – < 65	30
65 – < 85	9
totaal	150

Figuur 1.9 *Frequentieverdeling leeftijd*

We gaan in eerste instantie kijken hoe bij een frequentieverdeling de centrummaten worden berekend.

modale klasse

Bij een frequentieverdeling spreekt men eerder over de modale klasse, dan over de modus (zie figuur 1.4). Een brede klasse herbergt in principe meer waarnemingen dan een smalle klasse. Daarom wordt de modale klasse niet bepaald op basis van de frequentie, maar op basis van de frequentie per klassebreedte. Dit wordt de frequentiedichtheid genoemd. Een snelle manier om de frequentiedichtheden te berekenen is om alles te relateren aan een standaardklassebreedte. De standaardklassebreedte kan vrij worden gekozen. In ons voorbeeld zou dit de breedte van de klasse 10 tot < 15 kunnen zijn ofwel 5 jaar.

frequentiedichtheid

standaardklasse-breedte

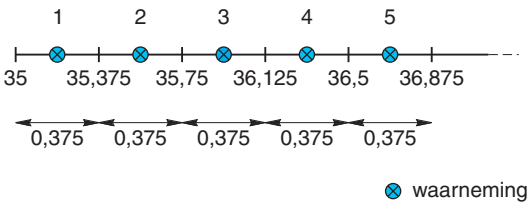
leeftijd	*frequentie*	*aantal maal standaardklassebreedte*	*frequentiedichtheid*
10 – < 15	10	1	10/1 = 10
15 – < 25	32	2	32/2 = 16
25 – < 35	29	2	29/2 = 14,5
35 – < 50	40	3	40/3 = 13,33
50 – < 65	30	3	30/3 = 10
65 – < 85	9	4	9/4 = 2,25
totaal	150		

Figuur 1.10 *Standaardklasse-breedte en frequentiedichtheid*

De modale klasse is nu: 15 tot < 25 jaar; deze klasse heeft de hoogste frequentiedichtheid.

De mediaan is de middelste waarneming. Van in totaal 150 waarnemingen zijn de 75ste en de 76ste waarneming de middelste waarnemingen. Dit zijn de vierde en de vijfde waarneming in de klasse 35 tot < 50.

Hoe de veertig waarnemingen in de klasse 35 tot < 50 jaar er precies uitzien is uit de frequentieverdeling niet op te maken. Men gaat er voor de eenvoud vanuit dat de waarnemingen gelijkmatig over de klasse verdeeld zijn.
De breedte van de klasse is vijftien jaar. Voor elke waarneming is dus een breedte van 15/40 = 0,375 jaar beschikbaar.
Het begin van de klasse 35 tot < 50 ziet er dan als volgt uit:



Figuur 1.11
De waarnemingen zijn gelijkmatig verdeeld

De eerste waarneming valt in het interval: 35 tot < 35,375, dus: 35,19 (afgerond).

VRAAG 1.15 Bereken de mediaan van de leeftijd.

De mediaan van de leeftijd is 36,5 jaar. Dit is het gemiddelde van de vierde en de vijfde waarneming van de leeftijdsklasse 35 tot < 50 jaar.

FORMULE 1.5 $Me = L + (r - \frac{1}{2}) * \frac{b}{f}$

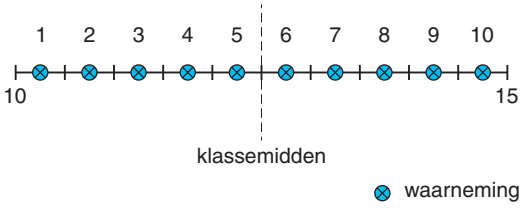
waarbij:
L: de linkergrens van de klasse waar de mediaan in valt;
r: het rangnummer binnen de klasse waar de mediaan in valt;
b: de breedte van de klasse waar de mediaan in valt;
f: het aantal waarnemingen van de klasse waar de mediaan in valt.

Voor 150 waarnemingen is de middelste: $\frac{1 + 150}{2} = 75{,}5$

Dit is de 4,5de waarneming binnen de klasse van 35 tot < 50. In bovenstaande formule kan voor r = 4,5 de mediaan direct worden berekend.

Bij het gemiddelde van een frequentieverdeling treedt weer het probleem op dat de 'losse' waarnemingen niet bekend zijn. Als we naar de eerste klasse van 10 tot < 15 jaar kijken, dan weten we alleen dat er tien waarnemingen in zijn gevallen. Of dit nu tien waarnemingen van tien jaar zijn of tien van 14,9 jaar is niet te achterhalen.

Evenals bij de bepaling van de mediaan gaan we ervan uit dat de waarnemingen gelijkmatig over de klasse verdeeld zijn.



Figuur 1.12
Klassemidden

klassemidden

In plaats van met alle tien individuele waarnemingen te gaan rekenen, is het handiger om deze samen te voegen tot één waarneming in het klassemidden met een gewicht van tien.
Het klassemidden kan bepaald worden door de laagste en de hoogste waarde binnen de klasse te middelen.

gewogen gemiddelde

Het gemiddelde kan worden bepaald als een gewogen gemiddelde van de klassemiddens. De eerste klasse van 10 tot < 15 jaar heeft een gewicht van 10, de tweede klasse van 15 tot < 25 jaar heeft een gewicht van 32, enzovoort.
In figuur 1.13 is de berekening van het gemiddelde uitgevoerd. De frequentie binnen een klasse wordt aangeduid met f_i. Het klassemidden van een klasse als m_i.

leeftijd	*frequentie* f_i	*klassemidden* m_i	$f_i m_i$
10 – < 15	10	12,5	125
15 – < 25	32	20	640
25 – < 35	29	30	870
35 – < 50	40	42,5	1 700
50 – < 65	30	57,5	1 725
65 – < 85	9	75	675
totaal	150		5 735

Figuur 1.13
Berekening gewogen gemiddelde

In totaal zijn de 150 personen 5 735 jaar oud. Het gemiddelde wordt verkregen door te delen door 150: $\frac{5\,735}{150} = 38{,}23$ jaar.

FORMULE 1.6 $\bar{x} = \dfrac{\sum f_i m_i}{n}$

Opmerking
Het is nauwkeuriger om met de 150 'losse' waarnemingen te rekenen, dan met de frequentieverdeling. Dit komt doordat er bij een frequentieverdeling aangenomen moet worden dat de waarnemingen binnen een klasse gelijkmatig verdeeld zijn, terwijl dit niet zo hoeft te zijn.

Bij een frequentieverdeling van een variabele uit de ratioschaal is het mogelijk de standaardafwijking te berekenen. De spreiding wordt veroorzaakt door de afstand tussen de klassemiddens en het gemiddelde. Deze afstanden moeten worden gekwadrateerd. Net als bij het gewogen gemiddelde moet rekening worden gehouden met het aantal waarnemingen binnen een klasse.
Figuur 1.14 geeft een overzicht van de berekening.

leeftijd	f_i	m_i	$m_i - x$	$(m_i - \bar{x})^2$	$f_i\,(m_i - \bar{x})^2$
10 – < 15	10	12,5	–25,73	662,03	6 620,30
15 – < 25	32	20	–18,23	332,33	10 634,56
25 – < 35	29	30	–8,23	67,73	1 964,17
35 – < 50	40	42,5	4,27	18,23	729,20
50 – < 65	30	57,5	19,27	371,33	11 139,90
65 – < 85	9	75	36,77	1 352,03	12 168,27
totaal	150			2 803,68	43 256,40

Figuur 1.14 *Berekening standaardafwijking*

VRAAG 1.16 Welke leeftijdsklasse (uit figuur 1.14) veroorzaakt de meeste spreiding? En welke de minste?

In totaal geeft dit voor alle waarnemingen samen: 43.256,40.

Ofwel: $\sum f_i(m_i - \bar{x})^2 = 43.256,40$

Deze waarde moet nog worden gemiddeld. Voor een steekproef moet dit worden gedeeld door het aantal waarnemingen minus één.

$$\frac{43.256,4}{150 - 1} = \frac{43.256,4}{149} = 290,31$$

Dit is de variantie s^2.

FORMULE 1.7

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

Als van de variantie de wortel wordt genomen dan wordt de standaardafwijking verkregen:

FORMULE 1.8

$$s = \sqrt{\frac{\sum f_i (m_i - \bar{x})^2}{n - 1}}$$

Dat geeft dus:

$$s = \sqrt{s^2} = \sqrt{290{,}3} = 17{,}04 \text{ jaar}$$

Bij de meeste rekenmachines kan deze berekening worden gedaan met de statistische 'mode'.

VOORBEELD 1.5

Casio fx-82SX Fraction

MODE		.		(zet de machine op statistiek)
SHIFT			AC	(maak het statistisch geheugen leeg)
12,5	X	10	M+	
20	X	32	M+	
30	X	29	M+	
42,5	X	40	M+	
57,5	X	30	M+	
75	X	9	M+	
SHIFT			6	(= 150 het aantal waarnemingen)
SHIFT			7	(= 38,23 het steekproefgemiddelde)
SHIFT			9	(= 17,04 de steekproefstandaardafwijking)
MODE			0	(zet de machine op rekenen)

De verdere resultaten (vraag 3 tot en met 6) van de volledige enquête onder 150 personen zijn weergegeven in de figuren 1.15 tot en met 1.18.

3 Van welke voorziening van de bibliotheek maakt u geregeld gebruik?

vraag 3 variabele	*nooit*	*1*	*2*	*3*	*4*	*5 altijd*
'infobalie'		33	44	44	26	3
'computerinfo'		29	23	26	46	26
'sanitair'		53	27	5	23	42
'fietsenstalling'		42	20	5	20	63

Figuur 1.15 *Gebruik voorzieningen*

4 Van welke afdeling maakt u geregeld gebruik?

vraag 4 variabele	*ja*	*nee*
'studiezaal'	44	106
'volwassenen'	116	34
'leeszaal'	28	122
'fonotheek'	39	111
'jeugdafdeling'	38	112

Figuur 1.16
Gebruik afdelingen

5 Hoeveel geld besteedt u maandelijks aan de koop van boeken, tijdschriften en kranten? (De resultaten zijn weergegeven in de vorm van een frequentieverdeling.)

variabele = 'geld'	*frequentie*
0 – < 5	45
5 – < 10	31
10 – < 15	25
15 – < 20	19
20 – < 30	15
30 – < 40	11
40 – < 60	3
60 – < 75	1
totaal	150

Figuur 1.17
Bestedingen aan boeken, tijdschriften en kranten

6 Wat is uw algemene oordeel over de bibliotheek in zijn geheel?

variabele = 'oordeel'	*frequentie*
zeer tevreden	48
tevreden	49
neutraal	26
ontevreden	20
zeer ontevreden	7
totaal	150

Figuur 1.18
Algemeen oordeel bibliotheek

VRAAG 1.17 Bereken de volgende centrummaten.

a Voor de variabele 'oordeel' (vraag 6).

S *b* Voor de variabele 'infobalie' (vraag 3).

Figuur 1.19 geeft een overzicht van alle behandelde centrummaten en spreidingsmaten.

centrummaten	*naam*
Mo	modus
Me	mediaan
$\bar{x}$	gemiddelde
spreidingsmaten	*naam*
R	spreidingsbreedte (range)
s	standaardafwijking (standaarddeviatie)
V	variatiecoëfficiënt

Figuur 1.19 *Overzicht centrummaten en spreidingsmaten*

1.1.3 Populatie en steekproef

Het voorgaande verhaal is in zijn geheel van toepassing op de situatie waarin de verzamelde gegevens een steekproef zijn van een groter geheel. Dit is de meest gebruikelijke situatie. In het geval dat de populatie geheel bekend is, moeten er kleine veranderingen in de formules worden aangebracht.

Uit de gegevens van de steekproef, van grootte n, wordt het steekproefgemiddelde ($\bar{x}$) en de steekproefstandaardafwijking (s) bepaald. Als de gegevens van de gehele populatie, van grootte N, bekend zijn, dan kan het populatiegemiddelde (μ) en de populatiestandaardafwijking (σ) worden berekend.

Voor 'losse' waarnemingen wordt dit:

FORMULE 1.9
$$\mu = \frac{\sum x_i}{N}$$

FORMULE 1.10
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Voor een frequentieverdeling wordt dit:

FORMULE 1.11
$$\mu = \frac{\sum f_i m_i}{N}$$

FORMULE 1.12

$$\sigma = \sqrt{\frac{\sum f_i(m_i - \mu)^2}{N}}$$

Het grootste verschil tussen een berekening met steekproefgegevens en populatiegegevens is dat bij het bepalen van de standaardafwijking bij populatiegegevens moet worden gedeeld door het aantal waarnemingen (N) en bij steekproefgegevens moet worden gedeeld door het aantal waarnemingen (n) minus één.

Op de rekenmachine is de populatiestandaardafwijking aanwezig.

VOORBEELD 1.6

Casio fx-82SX Fraction

SHIFT 8 (geeft de waarde van de populatiestandaardafwijking)

1.1.4 Grafieken

Een grafiek zegt vaak meer dan duizend woorden. Bij het maken van een correcte grafiek dient te worden gelet op de volgende onderdelen:

- Het opschrift: dit is een kernachtige weergave van de inhoud van de grafiek.
- Legenda: dit is een uitleg over de gebruikte kleuren en/of arceringen.
- Bronvermelding: indien bekend, moet worden vermeld waar de gegevens vandaan komen.
- Assen: geef aan wat de eenheid is van de as en geef de waarden aan op de as. Gebruik scheurlijntjes als de as niet bij nul begint.

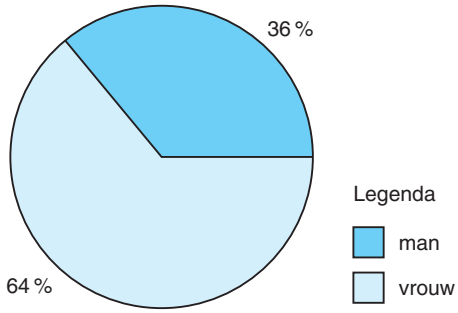
Op basis van de volledige steekproef van 150 personen kunnen voor de verschillende vragen diverse grafieken worden gemaakt. Voor de vragen 1 en 2 uit de enquête geven we een overzicht van grafieken die gebruikt kunnen worden om de gegevens te presenteren.

```
1  Van welk geslacht bent u?
   man    54
   vrouw  96
```

cirkeldiagram

Als de absolute aantallen er minder toe doen dan de verhouding tussen man en vrouw, is een cirkeldiagram geschikt (figuur 1.20).

geslacht	*aantal*	*percentage*	*graden*
man	54	36	129,6
vrouw	96	64	230,4
totaal	150	100	360

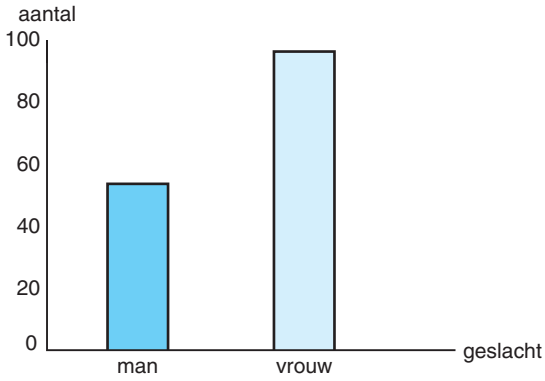


Figuur 1.20 *Cirkeldiagram* Bron: Bibliotheek Barometer 1996

Een cirkeldiagram is bruikbaar voor de weergave van alle soorten variabelen (nominaal, ordinaal en ratio).

staafdiagram

Als het wel van belang is om te zien hoe groot de aantallen zijn, is het mogelijk gebruik te maken van een staafdiagram. Eventueel kunnen op de verticale as percentages worden uitgezet.

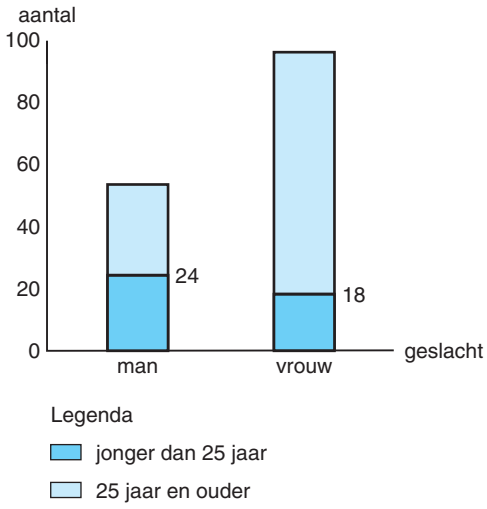


Figuur 1.21
Staafdiagram

Bron: Bibliotheek Barometer 1996

stapeldiagram

Stel dat bekend is dat van de groep 'man' er 24 jonger waren dan 25 jaar en dat er van de groep 'vrouw' achttien jonger waren dan 25 jaar. In het staafdiagram kan dit worden verwerkt door beide groepen op te splitsen. Zo'n diagram heet een stapeldiagram.



Figuur 1.22
Stapeldiagram

Bron: Bibliotheek Barometer 1996

Het is mogelijk om per staaf verschillende opsplitsingen te maken.

2 Hoe oud bent u?

leeftijd	*frequentie*
10 – < 15	10
15 – < 25	32
25 – < 35	29
35 – < 50	40
50 – < 65	30
65 – < 85	9
totaal	150

Figuur 1.23
Leeftijdsfrequenties

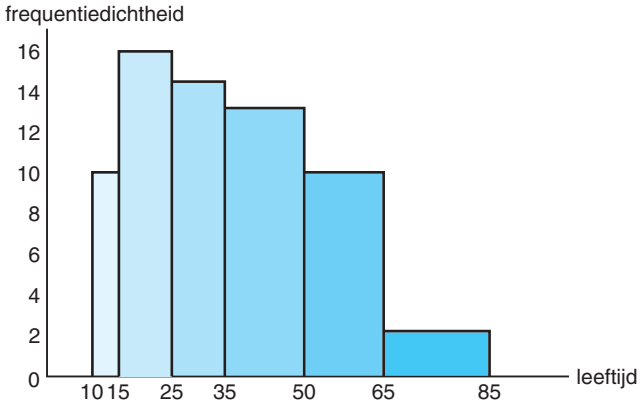
Het is niet gewenst om alleen naar de frequenties te kijken, omdat in een bredere klasse automatisch meer waarnemingen vallen. De frequenties moeten worden gedeeld door de bijbehorende klassebreedten. In paragraaf 1.1.2 *Frequentieverdeling* is dat al gedaan in figuur 1.10, bij de berekening van de modale klasse (met behulp van de standaardklassebreedte).

leeftijd	*frequentiedichtheid*
10 – < 15	10
15 – < 25	16
25 – < 35	14,50
35 – < 50	13,33
50 – < 65	10
65 – < 85	2,25

Figuur 1.24
Frequentiedichtheid

histogram

Als de frequentiedichtheden worden uitgezet ontstaat een histogram.



Figuur 1.25
Histogram

Bron: Bibliotheek Barometer 1996

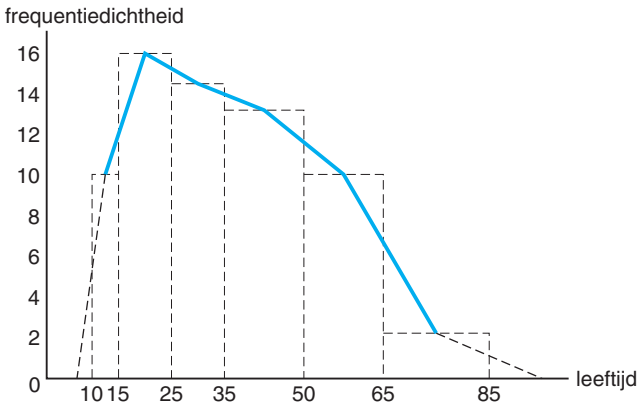
VRAAG 1.18 Welke centrummaat is direct uit het histogram af te lezen?

Typerend voor een histogram is dat de oppervlakte van een blokje in verhouding staat tot het aantal waarnemingen binnen een klasse.

Opmerking
Het is ook mogelijk om van 'losse' waarnemingen, van een ratioschaal, een histogram te maken. Daarvoor moeten wel vrij veel waarnemingen beschikbaar zijn. De waarnemingen moeten worden ingedeeld in klassen, zodat een frequentieverdeling ontstaat. Daarna gaat de procedure op dezelfde wijze als zojuist is beschreven. Als vuistregel mag worden aangenomen dat het aantal klassen ongeveer gelijk moet zijn aan de wortel van het aantal waarnemingen om een 'mooi' histogram te verkrijgen.

frequentiepolygoon

Een alternatief voor het gebruik van een histogram is het frequentiepolygoon. Hiervoor moeten de middens van de blokjes van het histogram (zie figuur 1.25) door rechte lijnen worden verbonden.
Aan de zijkanten van het frequentiepolygoon kunnen denkbeeldige klassen worden geconstrueerd om het polygoon naar nul door te trekken. Hiervoor zijn geen vastomlijnde regels bekend.
Een methode is om aan de linkerkant een klasse te construeren van dezelfde breedte als de eerste klasse (10 tot < 15 jaar) en het polygoon door te trekken naar het (denkbeeldige) klassemidden (7,5). Aan de rechterkant kan hetzelfde worden gedaan. Maak een extra klasse van 85 tot < 105 jaar en trek het polygoon door tot het klassemidden (95).



Figuur 1.26
Frequentiepolygoon

Bron: Bibliotheek Barometer 1996

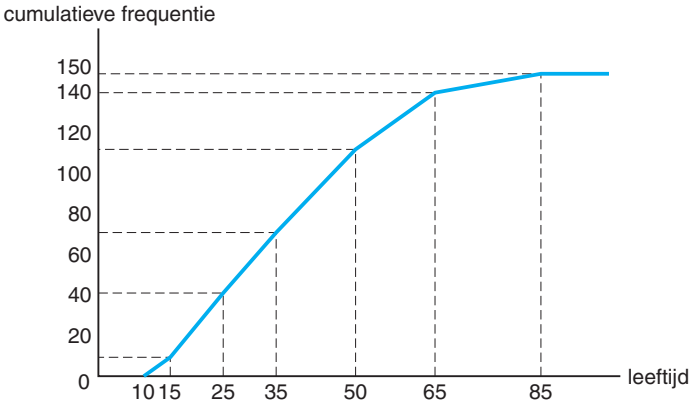
In het voorgaande is al eens een voorbeeld van een frequentiepolygoon afgebeeld. Kijk nog eens naar figuur 1.4 van het inkomen van vrouwen met een voltijdbaan.
Soms ben je niet alleen geïnteresseerd in de aantallen in een bepaalde klasse, maar ook in de aantallen onder een bepaalde grens. Bijvoorbeeld: het aantal personen jonger dan 25 is: 10 + 32 = 42.
Dit aantal (42) wordt een cumulatieve frequentie genoemd (denk aan cumulus = stapelwolk).

leeftijd	*frequentiedichtheid*	*cumulatieve frequentie*
10 – < 15	10	10
15 – < 25	32	42
25 – < 35	29	71
35 – < 50	40	111
50 – < 65	30	141
65 – < 85	9	150
totaal	150	

Figuur 1.27 *Cumulatieve frequentie*

cumulatief frequentiepolygoon

Als hiervan een grafiek wordt uitgezet met als verticale as de cumulatieve frequentie, dan ontstaat het cumulatief frequentiepolygoon.



Figuur 1.28 *Cumulatief frequentiepolygoon*

Bron: Bibliotheek Barometer 1996

VRAAG 1.19 Wat is het belangrijkste verschil bij het tekenen van een frequentiepolygoon en een cumulatief frequentiepolygoon als we het verschil wel of niet cumulatief even buiten beschouwing laten?

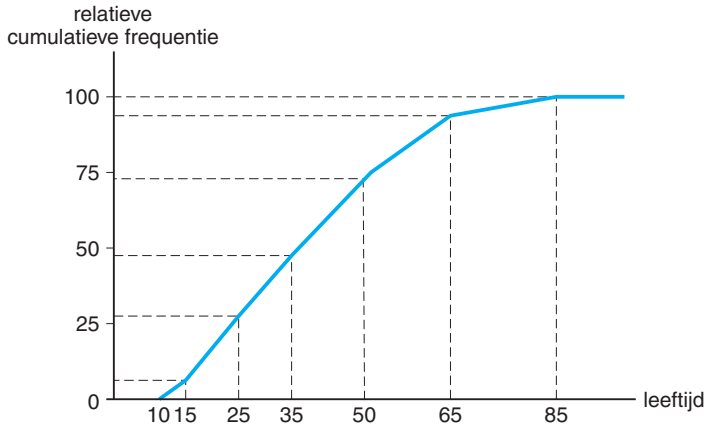
Als de aantallen niet van belang zijn, maar juist het percentage, dan kunnen de cumulatieve frequenties worden omgerekend in een percentage van het totaal (150) en ontstaan de relatieve cumulatieve frequenties.

leeftijd	*frequentie*	*cumulatieve frequentie*	*relatieve cumulatieve frequentie*
10 – < 15	10	10	6,7
15 – < 25	32	42	28,0
25 – < 35	29	71	47,3
35 – < 50	40	111	74,0
50 – < 65	30	141	94,0
65 – < 85	9	150	100,0
totaal	150		

Figuur 1.29 *Relatieve cumulatieve frequentie*

relatief cumulatief frequentiepolygoon

Als de grafiek wordt getekend met als verticale as de relatieve cumulatieve frequenties, dan ontstaat het relatief cumulatief frequentiepolygoon.

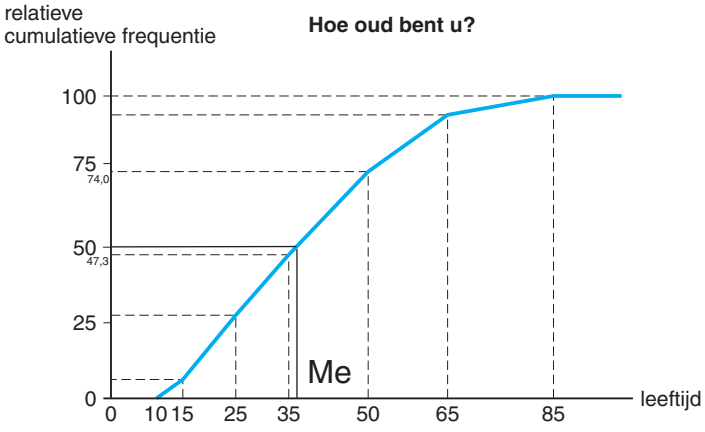


Figuur 1.30 *Relatief cumulatief frequentiepolygoon*

Bron: Bibliotheek Barometer 1996

VRAAG 1.20 Welke centrummaat kan uit het relatief cumulatief frequentiepolygoon worden afgelezen?

Het relatief cumulatief frequentiepolygoon kan (volgens de uitwerking van vraag 1.20) worden gebruikt om de mediaan te berekenen. De mediaan kan grafisch worden afgelezen uit de grafiek, maar kan ook worden berekend met de eerder gegeven formule 1.5.



Figuur 1.31
Aflezen mediaan

Bron: Bibliotheek Barometer 1996

Interpoleren geeft: $35 + \frac{50 - 47{,}3}{74 - 47{,}3} * (50 - 35) = 36{,}50$ jaar

Dit komt overeen met de uitwerking van vraag 1.15.

Kernbegrippen

aselect
cirkeldiagram
cumulatief frequentiepolygoon
frequentie
frequentiedichtheid
frequentiepolygoon
frequentieverdeling
gemiddelde
gewogen gemiddelde
histogram
klassemidden
mediaan
modus
modale klasse
nominale schaal
ordinale schaal
populatie
range
ratioschaal
relatief cumulatief frequentiepolygoon
spreidingsbreedte
staafdiagram
standaardafwijking
standaarddeviatie
standaardklassebreedte
stapeldiagram
steekproef
variantie
variatiecoëfficiënt

Formules

Steekproef

Losse waarnemingen

1.1 $\bar{x} = \frac{\sum x_i}{n}$ steekproefgemiddelde

1.2 $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$ steekproefvariantie

1.3 $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$ steekproefstandaardafwijking

1.4 $V = \frac{s}{\bar{x}}$ variatiecoëfficiënt bij steekproef

Frequentieverdeling

1.5 $Me = L + (r - \frac{1}{2}) * \frac{b}{f}$ mediaan

1.6 $\bar{x} = \frac{\sum f_i m_i}{n}$ steekproefgemiddelde

1.7 $s^2 = \dfrac{\sum f_i (m_i - \bar{x})^2}{n - 1}$ steekproefvariantie

1.8 $s = \sqrt{\dfrac{\sum f_i (m_i - \bar{x})^2}{n - 1}}$ steekproefstandaardafwijking

Populatie

Losse waarnemingen

1.9 $\mu = \dfrac{\sum x_i}{N}$ populatiegemiddelde

1.10 $\sigma = \sqrt{\dfrac{\sum (x_i - \mu)^2}{N}}$ populatiestandaardafwijking

$V = \dfrac{\sigma}{\mu}$ variatiecoëfficiënt bij populatie

Frequentieverdeling

1.11 $\mu = \dfrac{\sum f_i m_i}{N}$ populatiegemiddelde

1.12 $\sigma = \sqrt{\dfrac{\sum (f_i (m_i - \mu)^2}{N}}$ populatiestandaardafwijking

Opgaven

Maak nu ter afsluiting van dit hoofdstuk de opgaven in het toepassingenboek.