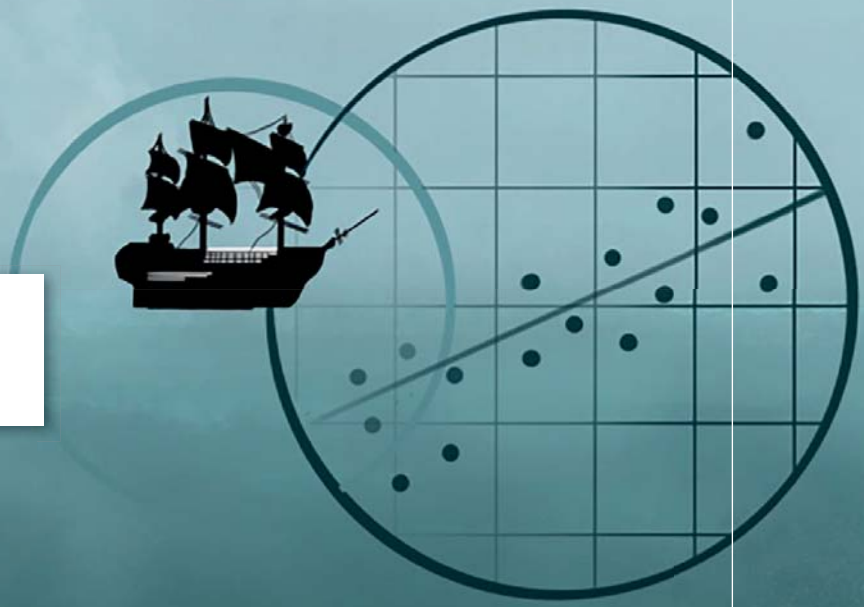
Statistiek in Stijl 1

Piraten, Perziken en P-waarden



Noordhoff

Vince Penders

1e druk

Statistiek in stijl 1

Piraten, perziken en p-waarden

Vince Penders

Eerste druk

Noordhoff Groningen/Utrecht

Ontwerp omslag: Shootmedia, Groningen
Omslagillustratie: Chelsy Penders

Eventuele op- en aanmerkingen over deze of andere uitgaven kunt u richten aan: Noordhoff Uitgevers bv, Afdeling Hoger Onderwijs, Antwoordnummer 13, 9700 VB Groningen of via het contactformulier op www.mijnnoordhoff.nl.

0 / 21

ISBN (ebook) 978-90-01-74686-5
ISBN 978-90-01-74685-8
NUR 916

Zij die gemaakt hebben, groeten u

Welkom, beste lezer! Voor je neus ligt een splinternieuw handboek, en wat zijn we trots op het resultaat. Als makers willen wij ons kort aan je voorstellen. Mag dat?

Die jongeman links is **Vince Penders** ... noem me maar **Vincenzo**. Het brein achter deze lesmethode – dat ben ik. Creatief tot op het bot, een vrolijke perfectionist. Vroeger was ik werkzaam bij de Universiteit Maastricht, maar tegenwoordig werk ik als onafhankelijk docent en geef jaarlijks les aan grofweg 300 studenten. In de tussentijd verbeter ik constant mijn lesboeken. Op de zeldzame momenten waarop ik niet werk of onder vrienden ben, sta ik te koken, speel ik een Nintendospel of schrijf ik fictie met epische muziek in mijn oren. Jawel: verhalen zijn mijn tweede (of eigenlijk eerste) passie. Mijn meest recente roman, *Een beer en een boerka*, vertelt het relaas van een man die geloofde dat hij iets kleins en pluizigs was. Evenals mijn debuut *Zwaluwhart* is het uitgebracht door Macc.

Mijn zus **Chelsy Penders** (vooraan) zingt, danst en werkt als interieurstylist. Waar mijn artistieke stem het geschreven woord is, is de hare de handgemaakte tekening in elke vorm en stijl. Chelsy's veelzijdigheid blijkt wel uit de vele illustraties die dit boek rijk is, evenals uit de gelikte kleuren-

schema's (vergeef me mijn gebrek aan trots op haar werk). Haar Maine Coon-katten worden graag knuffelig als ze thuis zit te tekenen, dus laat het ons weten als je ergens een kattenpoot tegenkomt.

En dit is **Sophia von Stockert**, een oud-student van me en tegenwoordig een goede vriendin. Als toegewijd statistiekdocent inspireerde ze vroeger studenten om te ontdekken wat voor coole dingen je allemaal met statistiek en onderzoek kunt doen. Om deze reden heb ik haar aan boord gehaald als redacteur van *Statistiek in stijl*. Sophia heeft een diploma in de Research Master Cognitive and Clinical Neuroscience en volgt momenteel een opleiding tot psychotherapeut in Heidelberg. In haar vrije tijd klimt ze graag en ze redigeert en vertaalt boeken en manuscripten.

Dan nog een paar eervolle vermeldingen als dankbetuigingen. De eerste is voor Olav Wessel, die het lef had om mijn manuscript telefonisch bij Noordhoff aan te bieden en ons vervolgens wist binnen te loodsen via een obscuur online formulier van de klantenservice. De tweede gaat naar Vincent Diks, mijn uitgever, die vanaf het eerste moment in dit boek geloofde. Ten slotte een dikke omhelzing voor alle studenten die me elke dag inspireren. Jullie helpen is een eer. Jullie hebben me ertoe overgehaald dit boek te schrijven en mijn passie brandende gehouden om het verder te verbeteren met deze nieuwe editie. Mensen bedanken me vaak voor het werk dat ik doe, maar er bestaat slechts één juist antwoord. 'Nee… *jullie* bedankt!'

Inhoud

Hoe gebruik je dit boek?

Statistiek in stijl kan fungeren als een op zichzelf staande cursus statistiek. In de praktijk zullen de meeste lezers het waarschijnlijk gebruiken ter ondersteuning van een vak op de hogeschool of universiteit. Daarom hebben we de cursus opgesplitst in twee aparte boeken:

- *Statistiek in stijl 1* dekt de basis af. Steekproeven beschrijven, kanstheorie, hypothesetoetsen, betrouwbaarheidsintervallen … Waarschijnlijk heb je iets aan dit deel als je een hogeschoolstudent bent of in je eerste jaar van de universiteit zit.
- *Statistiek in stijl 2* opent het volledige arsenaal. Statistische modellen met meer dan twee variabelen, herhaalde metingen, psychometrie … Zet koers naar deze wateren in je tweede jaar van de universiteit en verder.

Uiteraard is dit slechts een richtlijn. Individuele curricula bespreken mogelijk wat onderwerpen uit boek 2 op de hogeschool of komen terug op stof van boek 1 gedurende het tweede universiteitsjaar. Onderzoekers (en perfectionistische studenten) kunnen het als prettig ervaren om beide boeken te bezitten, zodat ze de kleine details kunnen nalezen. Merk op dat een bezitter van enkel boek 2 zich geen zorgen hoeft te maken: allerlei sleutelbegrippen uit boek 1 staan samengevat in de sectie *Overzichten en kernbegrippen*! Gebruik deze voor een soepele studeerervaring. ☺
Elk hoofdstuk kent zijn eigen moeilijkheidsgraad en vereist soms voorkennis uit eerdere hoofdstukken. Dit staat telkens aangegeven op de introductiepagina. De moeilijkheidsgraden zijn als volgt:

Niveau	Toelichting
 Cupcake	Voor de nieuwkomer in de wereld van de statistiek. Geen wiskundeknobbel? Geen probleem! Ook jij gaat na een dag studeren met een gerust hart naar bed.
 Zeemeeuw	Hier wordt het wat spannender. Geduld en herhaling zijn de ingrediënten waarmee je de eindstreep haalt. Moeilijker dan dit gaan we het in boek 1 niet maken.
 Gorilla	Het echte werk. Veel wetenschappelijk onderzoek is op dit soort statistiek gebaseerd. De meeste hoofdstukken in boek 2 zijn van gorillaniveau.
 Ninja	Expert in de dop, hier moet je wezen. Neem je je vak serieus, dan leren deze hoofdstukken je behoorlijk geavanceerde technieken. Voorkennis en motivatie zijn een must. In ruil daarvoor wordt statistiek cooler dan ooit tevoren.

Behalve theoretische uitleg bevat elk hoofdstuk ook *opdrachten*. Ik heb deze zo ontworpen dat ze je kennis niet slechts op de proef stellen, maar ook verdiepen. Oefenen dus! Goede opdrachten zijn leerzaam en voegen veel toe aan je praktische ervaring. In de meeste hoofdstukken vind je maar liefst twee setjes. De *opdrachten voor perziken* zijn het fijnst om mee te beginnen; de *opdrachten voor piraten* zijn wat uitdagender.
Oh, en de uitwerkingen? Die kun je gratis downloaden van de website *statistiekinstijl.noordhoff.nl*. Op deze site zie je overigens ook meteen alle kanstabellen.

Ben je zelfstandig bezig met statistiek?
Begin dan gewoon bij hoofdstuk 1 of het onderwerp waarover je graag iets wilt leren. Zie het volgende kopje voor bijzondere afwegingen.

Volg je een statistiekvak op de hogeschool of universiteit?
Ga naar de inhoudsopgave en zoek het onderwerp op dat je moet leren. Kun je het niet vinden, raadpleeg dan eens het register achter in het boek. De kans bestaat dat jouw onderwerp besproken wordt in *boek 2*!
Gevonden? De voorkennis die je nodig hebt, staat telkens aan het begin van het hoofdstuk vermeld. Lees eerst met een goede kop koffie of thee de theorie en maak aantekeningen. Schrijf een eigen samenvatting, vergelijk deze op het eind met die van mij en verbeter je eigen exemplaar waar nodig. Maak daarna de opdrachten om je nieuwverworven kennis op de proef te stellen. De uitwerkingen vind je gratis op de website.
Raak vooral niet ontmoedigd als de opdrachten je niet meteen goed afgaan. Statistiek heeft aandacht en oefening nodig. Soms gaat het een dag later al veel beter, als de nieuwe stof een beetje is bezonken. Veel succes! ☺

Heb je statistiek nodig voor je eigen onderzoek?
Waarschijnlijk ben jij al bekend met de onderwerpen die ik in dit boek bespreek. Kies de juiste statistische analyse met behulp van het stroomschema in de appendix, en ga dan naar het bijbehorende hoofdstuk.
In de online leeromgeving vind je ook instructies om Excel of SPSS aan te sturen.

DEEL 1
Data beschrijven

In de kwantitatieve wetenschap speelt het verzamelen en analyseren van gegevens een belangrijke rol. Deze eerste module leert je die gegevens optimaal te organiseren en te beschrijven. Tel ijsjes, bezoek een spookhuis en breng wat pit in je romantische diner; een goed begin is het halve werk. Er wacht een nieuwe wereld op je!

1 Gegevens in kaart

Over zomerse verkoeling en z-transformaties

Niveau
Cupcake
Vereiste voorkennis
Nihil (dit is het eerste hoofdstuk, weet je ☺)
Leerdoelen
Na dit hoofdstuk kun je Soorten variabelen onderscheiden Het gedrag van variabelen samenvatten in woord en beeld, en met kengetallen Variabelen omzetten naar een andere meeteenheid Het bovenstaande voor elkaar krijgen met behulp van een computer

1.1 Zoveel smaken: variabelen

Dit eerste hoofdstuk vertelt je hoe we data overzichtelijk kunnen presenteren, samenvatten wat er in een dataset gebeurt, en opvallende waarden eruit vissen. Klinkt eng en abstract? Ah, beste lezer, iedereen vergist zich wel eens: het wordt juist hartstikke leuk. Laten we het meteen concreet maken. De ijsjes vliegen over de toonbank in ijssalon Peach's Castle. Met name in het hoogseizoen zijn de klanten nauwelijks bij te houden. We gaan een steekproef trekken van 72 ijsjes die vandaag zijn verkocht om een beeld te krijgen van de populairste smaken en soorten – handig voor de salon. Een grotere steekproef zou natuurlijk nog representatiever zijn, maar als nieuwe lezer krijg je deze opzet wat gemakkelijker door je keel. De volgende dingen heb ik genoteerd:

- Welke smaak heeft het ijsje? Peach's Castle verkoopt maar liefst 10 verschillende smaken, waaronder truffelchocolade, karamel met nootjes en perzikparade (de specialiteit).
- In welk hoorntje wordt het ijs geserveerd: een minibakje, een oubliehoorn of een reuzenhoorn?
- Bij welke temperatuur (in graden Celsius) wordt het geserveerd?
- Hoeveel bolletjes koopt de klant?

Variabele

Deze eigenschappen zullen per klant verschillen. Daarom noemen we het aantal bolletjes, de smaak enzovoort *variabelen*: grootheden waarvan de waarde varieert, in dit geval afhankelijk van de deelnemer.

Meetniveau

De genoemde variabelen zijn niet van dezelfde aard. Elke heeft zijn eigen *meetniveau*:

I Nominaal

Categorische variabele

Dit meetniveau behoort tot het algemene type van *categorische* (of kwalitatieve) variabelen en staat voor 'benoeming'. Welke smaak heeft het ijsje? Er zijn tien verschillende antwoorden mogelijk, waaronder 'perzikparade' en 'latte macchiato'. In deze antwoorden zit geen rangorde; het zijn *gelijkwaardige categorieën*. We kunnen er wel een nummer aan toekennen, maar dit nummer betekent op zichzelf niets.

Nominale variabelen die je vaak tegenkomt in sociaalwetenschappelijk onderzoek zijn binair: geslacht (man of vrouw), nationaliteit (Nederlands, Duits, Frans et cetera) en bijvoorbeeld religie (geen, christelijk, islamitisch en ga zo maar door).

Dichotome, trichotome of polytome variabele

Soms worden ze trouwens aangeduid op basis van het aantal categorieën dat ze hebben: *dichotoom* (2; 'tweewaardig'), *trichotoom* (3; 'driewaardig') of *polytoom* (meer dan 3; 'veelwaardig').

II Ordinaal

Ook dit meetniveau is nog categorisch en staat voor 'ordening'. In wat voor een hoorntje wordt het ijs geserveerd? De mogelijke antwoorden zijn 'minibakje', 'oubliehoorn' en 'reuzenhoorn'. Deze categorieën kennen een duidelijke *rangorde*, want de hoorn wordt steeds groter. We zouden er nummers aan kunnen toekennen, zoals 1, 2, 3. Hoe hoger het nummer, hoe hoger de rang van het hoorntje (de grootte dus). Maar het verschil tussen 1 en 2 (minibakje versus oubliehoorn) is niet per se even groot als

het verschil tussen 2 en 3 (oubliehoorn versus reuzenhoorn). Daarom betekenen de nummers nog steeds niets op zichzelf.
Typische ordinale variabelen in sociaalwetenschappelijk onderzoek zijn hoogst genoten opleiding (vaak gegroepeerd als laag, midden en hoog), inkomensgroep (lagere, midden- en hogere klasse), en werknemersstatus (junior, manager, directeur).

III Interval

Kwantitatieve variabele

Dit meetniveau is het eerste van de *kwantitatieve* niveaus. Bij welke temperatuur wordt het ijs geserveerd? Het antwoord meten we in graden Celsius en het verschil tussen 0 en 1 graden is even groot als het verschil tussen 3 en 4 graden. Deze constante verschillen of afstanden voorzien de getallen van een ware kwantitatieve betekenis: je kunt ermee rekenen. Vandaar ook het etiket 'interval', dat 'afstand tussen twee punten' betekent. Sociale wetenschappers meten heel wat persoonlijkheidskenmerken die we niet direct kunnen 'zien', zoals intelligentie en religiositeit; hiervoor gebruiken ze tests en vragenlijsten. De abstracte testscore is vaak een intervalschaal. Denk aan IQ (50-150).

IV Ratio

Ook een kwantitatieve variant. Hoeveel bolletjes ijs koopt de klant? Dat is een intervalvariabele met een *absoluut nulpunt*: het is mogelijk om 0 bolletjes te kopen (door de ijssalon te passeren), maar minder dan nul scoren is onmogelijk. Dat maakt het zinvol om over *verhoudingen* (ratio's) te praten: vier bolletjes zijn twee keer zoveel als twee, bijvoorbeeld. Deze vlieger gaat niet op voor intervalvariabelen; je kunt niet zeggen dat 6 graden Celsius twee keer zo warm is als 3 (juist omdat je ook onder nul kunt gaan). Voor de statistiek maakt het onderscheid tussen interval- en ratiovariabelen meestal niets uit. Tot andere voorbeelden uit de sociale wetenschap behoren veel natuurlijke verschijnselen, zoals leeftijd, lichaamsgewicht en reactietijd.

Het meetniveau van een variabele bepaalt hoe we de scores op die variabele kunnen samenvatten en afbeelden. In latere hoofdstukken zal het meetniveau bepalen welke statistische toetsen we op een variabele kunnen loslaten. Meetniveaus zijn dus zowel gemakkelijk als cruciaal.

Zijn er nog vragen?

Vincenzo, als voorbeeld van een nominale variabele noemde je zojuist 'geslacht (man of vrouw)'. Vind je dat niet een beetje ouderwets?
Leuke vraag, beste lezer! Dankzij de maatschappelijke discussie over gender zien steeds meer mensen geslacht tegenwoordig als een intervalspectrum. Maar honderd jaar geleden was 'geslacht' waarschijnlijk een ordinale variabele … Snap je 'm? Zo zie je dat het meetniveau van sommige variabelen niet bepaald is door de God van de Wiskunde, maar afhangt van onze cultuur en hoe we tegen de wereld aankijken. Is dat niet fraai?

1.2 Het ontwerp van het menu: tabellen en grafieken

Terug naar het onderzoek: we hebben in totaal 72 bestellingen gevolgd. Tabel 1.1 presenteert de gegevens. Opmerking: één klant stelde het niet op prijs dat we een thermometer in zijn ijsje wilden steken om de temperatuur te meten. Daardoor bleef de temperatuur van dit ijsje helaas onbekend. Het incident maakte hem boos, dus hij liet ons ook niet zijn hoorntje zien.

TABEL 1.1 De gegevens van 72 klanten van Peach's Castle

Klantnummer	Smaak	Hoorntje	Temperatuur	Bolletjes
1	perzikparade	reuzenhoorn	−0,4	3
2	vanille voor dummies	minibakje	−2,2	1
3	vanille voor dummies	???	???	4
4	perzikparade	oubliehoorn	0,4	2
5	perzikparade	oubliehoorn	−1,9	2
6	perzikparade	minibakje	−1,4	2
7	truffelchocolade	oubliehoorn	−2,8	3
8	perzikparade	oubliehoorn	−2,5	2
9	perzikparade	reuzenhoorn	−0,3	8
10	vanille voor dummies	oubliehoorn	−1,4	3
11	bosvruchtenmelange	minibakje	−1,3	1
12	laurierdrop	minibakje	−0,5	1
13	karamel met nootjes	reuzenhoorn	−3,0	3
14	perzikparade	reuzenhoorn	−1,6	4
15	citroen en basilicum	reuzenhoorn	−1,1	4
16	laurierdrop	reuzenhoorn	−1,0	3
17	perzikparade	minibakje	0,0	1
18	citroen en basilicum	oubliehoorn	−2,6	2
19	bosvruchtenmelange	oubliehoorn	−0,8	2
20	kersen met koekjes	minibakje	−1,2	1
21	vanille voor dummies	oubliehoorn	−1,4	2
22	truffelchocolade	oubliehoorn	0,1	3
23	vanille voor dummies	oubliehoorn	−1,5	2
24	kersen met koekjes	reuzenhoorn	−1,9	4
25	citroen en basilicum	minibakje	−2,9	1
26	vanille voor dummies	reuzenhoorn	−1,1	3
27	groene appel	minibakje	−4,0	1

TABEL 1.1 De gegevens van 72 klanten van Peach's Castle (vervolg)

Klantnummer	Smaak	Hoorntje	Temperatuur	Bolletjes
28	perzikparade	reuzenhoorn	−1,0	5
29	bosvruchtenmelange	reuzenhoorn	−0,9	3
30	perzikparade	reuzenhoorn	−0,8	5
31	perzikparade	oubliehoorn	−2,2	2
32	truffelchocolade	oubliehoorn	−1,6	3
33	groene appel	oubliehoorn	−2,6	2
34	vanille voor dummies	oubliehoorn	−2,7	2
35	citroen en basilicum	minibakje	−3,2	1
36	citroen en basilicum	minibakje	−2,2	1
37	karamel met nootjes	reuzenhoorn	−1,9	4
38	latte macchiato	minibakje	−1,3	1
39	truffelchocolade	oubliehoorn	−0,7	2
40	perzikparade	oubliehoorn	−0,4	2
41	truffelchocolade	minibakje	−1,9	1
42	laurierdrop	reuzenhoorn	−2,6	3
43	karamel met nootjes	minibakje	−0,6	1
44	perzikparade	oubliehoorn	−0,4	2
45	perzikparade	oubliehoorn	−2,9	2
46	groene appel	minibakje	−0,7	1
47	karamel met nootjes	oubliehoorn	−1,5	2
48	karamel met nootjes	oubliehoorn	−1,7	3
49	perzikparade	oubliehoorn	−1,1	2
50	truffelchocolade	minibakje	0,0	1
51	kersen met koekjes	minibakje	−2,8	1
52	vanille voor dummies	oubliehoorn	0,0	2
53	bosvruchtenmelange	minibakje	−0,2	1
54	laurierdrop	reuzenhoorn	0,6	3
55	kersen met koekjes	reuzenhoorn	−0,8	3
56	citroen en basilicum	reuzenhoorn	−1,9	4
57	perzikparade	oubliehoorn	−0,4	1
58	truffelchocolade	oubliehoorn	1,0	2
59	vanille voor dummies	minibakje	−3,3	1
60	latte macchiato	oubliehoorn	−0,8	2
61	karamel met nootjes	oubliehoorn	−2,6	2
62	perzikparade	oubliehoorn	−2,2	2

TABEL 1.1 De gegevens van 72 klanten van Peach's Castle (vervolg)

Klantnummer	Smaak	Hoorntje	Temperatuur	Bolletjes
63	truffelchocolade	oubliehoorn	−1,4	3
64	kersen met koekjes	reuzenhoorn	−1,3	4
65	citroen en basilicum	oubliehoorn	−1,7	2
66	karamel met nootjes	reuzenhoorn	−2,7	4
67	laurierdrop	oubliehoorn	−2,0	2
68	latte macchiato	minibakje	−0,2	1
69	citroen en basilicum	minibakje	−1,3	1
70	kersen met koekjes	oubliehoorn	0,8	2
71	perzikparade	minibakje	−1,8	1
72	truffelchocolade	oubliehoorn	−2,5	2

Zo begin je meestal als onderzoeker … met een onoverzichtelijke waslijst. ☹ Gelukkig heeft de mensheid in de loop der eeuwen heel wat manieren bedacht om deze gegevens duidelijk (en liefst ook aantrekkelijk) weer te geven. In deze paragraaf nemen we een flinke greep uit het assortiment. Volg je een statistiekcursus, vraag dan aan je docent wat je moet kennen.

Univariate frequentieverdeling

Alle tabellen en grafieken in deze paragraaf heten trouwens *univariate frequentieverdelingen*: het zijn verdelingen van één *(uni)* variabele tegelijk. Wordt vervolgd, want twee of meer variabelen kunnen ook een verband vertonen. Met die verbanden gaan we ons bezighouden in hoofdstuk 2 en 3.

I Frequentietabel

Nominaal | Ordinaal | Interval | Ratio

Deze tabel geeft van elke categorie of waarde aan hoe vaak ze gescoord is. Neem om te beginnen tabel 1.2. We zien dat van de 72 ijsjes (totaal) er 4 smaakten naar bosvruchten, 8 naar citroen en basilicum, 9 naar truffelchocolade enzovoort. Om vast te stellen welke smaken relatief populair waren, kunnen we elke frequentie ook omzetten naar een percentage. Dat gaat zo:

$$percentage = \frac{frequentie}{totaal} \times 100\%$$

De 18 ijsjes met perziksmaak, bijvoorbeeld, vormden $\frac{18}{72} \times 100\% = 25\%$ van het totaal. Oftewel, een kwart van de klanten (1 op de 4) heeft voor de specialiteit van Peach's Castle gekozen.

Vanaf ordinale variabelen is het ook zinvol om in de frequentietabel het cumulatieve ('opgestapelde') percentage erbij te zetten: zie tabel 1.3.

De 74,7% bij 'oubliehoorn' is de 29,6% van de voorgaande categorie, plus de 45,1% van 'oubliehoorn'. Dus als we de minibakjes en de oubliehoorntjes (de laagste twee rangen) bij elkaar nemen, hebben we 74,7% van de ijsjes. Tellen we ook de reuzenhoorns erbij op, dan hebben we een cumulatief percentage van 100. Dit kunnen we niet doen met de smaakvariabele, aangezien nominale categorieën geen vaste volgorde hebben; we zouden steeds andere cumulatieve percentages krijgen door willekeurig een nieuwe volgorde te hanteren.
Frequentietabel 1.4 ten slotte probeert de scores op een intervalvariabele samen te vatten. Tja … dat lukt niet echt. De temperatuur van de ijsjes kan zoveel verschillende waarden aannemen dat het aantal rijen in de tabel de pan uitrijst. Voor kwantitatieve variabelen kun je beter iets anders gebruiken.

TABEL 1.2 Frequentietabel van de smaakvariabele (nominaal). Werkt goed

		Frequentie	Percentage
Smaak	vanille voor dummies	9	12,5
	citroen en basilicum	8	11,1
	groene appel	3	4,2
	karamel met nootjes	7	9,7
	bosvruchtenmelange	4	5,6
	kersen met koekjes	6	8,3
	latte macchiato	3	4,2
	perzikparade	18	25,0
	laurierdrop	5	6,9
	truffelchocolade	9	12,5
	Totaal	**72**	**100,0**

TABEL 1.3 Frequentietabel van de hoorntjesvariabele. Werkt uitstekend vanwege het kleine aantal categorieën

		Frequentie	Percentage	Cumulatief percentage
Hoorntje	minibakje	21	29,6	29,6
	oubliehoorn	32	45,1	74,7
	reuzenhoorn	18	25,4	100,0
	Totaal	**71**	**100,0**	

TABEL 1.4 Frequentietabel van de temperatuurvariabele. Werkt voor geen meter

		Frequentie	Percentage	Cumulatief percentage
Temperatuur	−4,0	1	1,4	1,4
	−3,3	1	1,4	2,8
	−3,2	1	1,4	4,2
	−3,0	1	1,4	5,6
	−2,9	2	2,8	8,5
	−2,8	2	2,8	11,3
	−2,7	2	2,8	14,1
	−2,6	4	5,6	19,7
	−2,5	2	2,8	22,5
	−2,2	4	5,6	28,2
	−2,0	1	1,4	29,6
	−1,9	5	6,9	36,6
	−1,8	1	1,4	38,0
	−1,7	2	2,8	40,8
	−1,6	2	2,8	43,7
	−1,5	2	2,8	46,5
	−1,4	4	5,6	52,1
	−1,3	4	5,6	57,7
	−1,2	1	1,4	59,2
	−1,1	3	4,2	63,4
	−1,0	2	2,8	66,2
	−0,9	1	1,4	67,6
	−0,8	4	5,6	73,2
	−0,7	2	2,8	76,1
	−0,6	1	1,4	77,5
	−0,5	1	1,4	78,9
	−0,4	4	5,6	84,5
	−0,3	1	1,4	85,9
	−0,2	2	2,8	88,7
	0,0	3	4,2	93,0
	0,1	1	1,4	94,4
	0,4	1	1,4	95,8
	0,6	1	1,4	97,2
	0,8	1	1,4	98,6
	1,0	1	1,4	100,0
	Totaal	**71**	**100,0**	

II Stamdiagram

Nominaal Ordinaal **Interval** **Ratio**

Deze boomachtige tabel kunnen we maken als de scores (zoals ijstemperaturen) deels of allemaal uit meerdere cijfers bestaan. De linkerkolom heet de 'stam'; die bevat bijvoorbeeld tientallen. In de rechterkolom vinden we de 'bladeren' die eraan vasthangen. In het geval van tabel 1.5 staan op de stam de gehele getallen, en op de bladeren de decimalen. De eerste score is –4,|0 oftewel –4,0 graden. De tweede score is –3,|0 oftewel –3,0 graden. Zo kunnen we doorgaan: –3,2; –3,3; –2,0; –2,2 et cetera. Het stamdiagram is wat overzichtelijker dan de frequentietabel, omdat scores die op elkaar lijken samen op één blad worden gezet.

Tweezijdig stamdiagram

Wil je twee groepen apart bekijken, dan kun je een stamdiagram ook splitsen. We noemen tabel 1.6 een *tweezijdig stamdiagram*. De stam bevindt zich nu in het midden. De bladeren links geven de ijstemperaturen aan van de kinderen die een ijsje kochten; de bladeren rechts betreffen de ijstemperaturen van de volwassen klanten. Zoals je ziet, zijn de twee groepen ongeveer hetzelfde verdeeld over de temperatuurschaal; het is niet zo dat de ene groep om heel koude ijsjes heeft gevraagd en de andere groep om lauwe ijsjes. De klanten hadden hierin nu eenmaal niks te willen. Soms verraadt een stamdiagram echter een verschil tussen de twee groepen.

Werp tot slot eens een blik op het stamdiagram van de bolletjesvariabele (tabel 1.7). Het maakt een ietwat merkwaardige indruk, want we hebben alleen gehele getallen; het laat dan ook vooral zien welke aantallen bolletjes vaak gekozen worden, met name 2,0 (2) bolletjes. Stamdiagrammen zijn dus vooral nuttig als de waarden van je variabele uit minstens twee cijfers bestaan.

TABEL 1.5 Stamdiagram van de temperatuurvariabele

Temperatuur	
–4,	0
–3,	023
–2,	02222556666778899
–1,	00111233334444556677899999
–0,	2234444567788889
0,	0001468
1,	0

TABEL 1.6 Tweezijdig stamdiagram van de temperatuurvariabele, uitgesplitst naar leeftijdsgroep

Temperatuur		
kind		***volwassene***
0	–4,	
3	–3,	02
7766522	–2,	0225668899
9987765544320	–1,	0111333446999
9865444432	–0,	277888
81000	0,	46
0	1,	

TABEL 1.7 Stamdiagram van de bolletjesvariabele

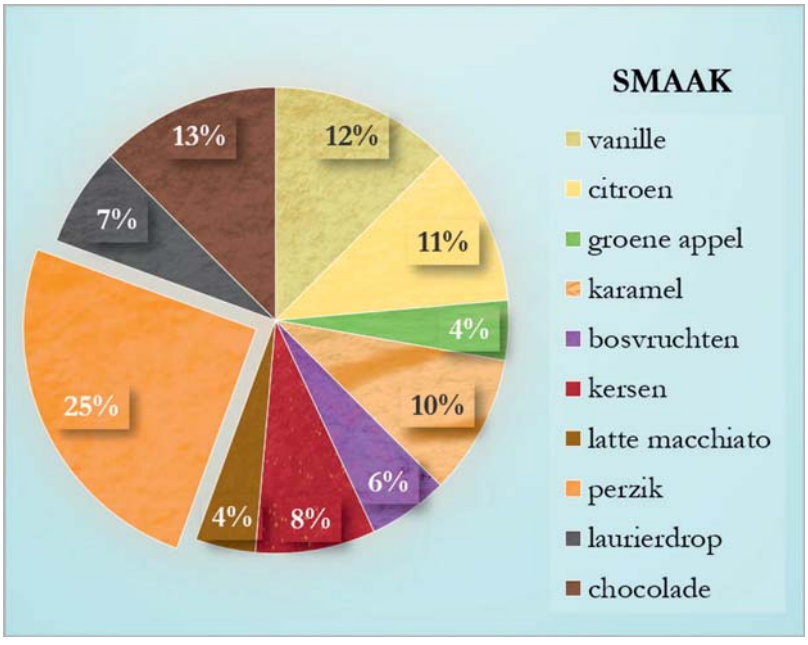
Bolletjes	
1,	00000000000000000000
2,	00000000000000000000000000
3,	000000000000000
4,	00000000
5,	00
6,	
7,	
8,	0

III Cirkeldiagram

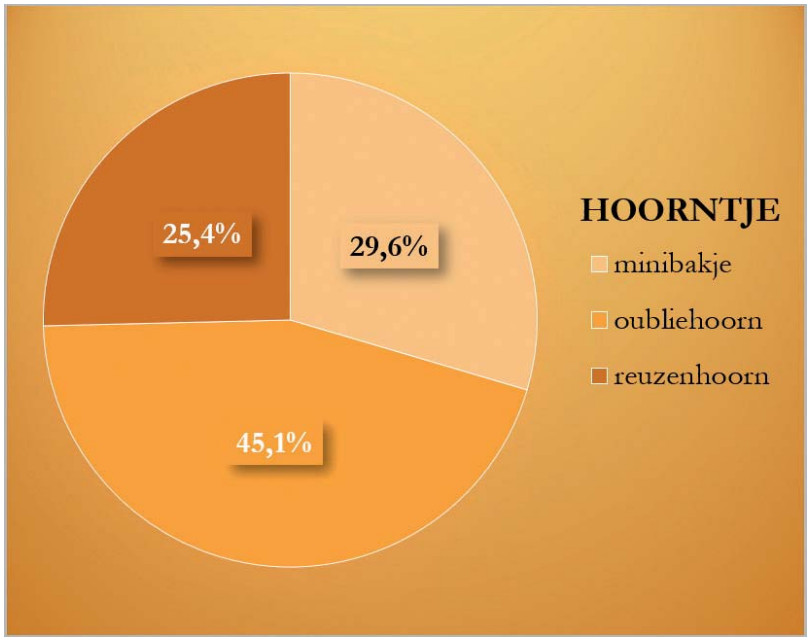
Nominaal Ordinaal Interval Ratio

Dit diagram verbeeldt de percentages uit de frequentietabel als taartpunten (de Engelsen noemen het dan ook een *pie chart*). Hoe groter het percentage, des te groter de punt. Desgewenst kun je ook de 'gewone' frequenties aangeven. Wil je een bepaalde categorie benadrukken, zoals de perzikparadesmaak, dan kun je de desbetreffende taartpunt een stukje uit de cirkel laten springen. Je zou er spontaan trek in ijstaart van krijgen … Cirkeldiagrammen kun je het best gebruiken als je iets nominaals hebt gegeten, eh, gemeten. Voor ordinale variabelen vormen ze niet zo'n toffe grafiek, omdat we de rangorde er niet in kunnen aangeven: de cirkel is rond, dus wat is de hoogste rang en wat de laagste? Tip: probeer eens een staafdiagram.

FIGUUR 1.1 Cirkeldiagram van de smaakvariabele. Werkt!



FIGUUR 1.2 Cirkeldiagram van de hoorntjesvariabele. Kan beter.

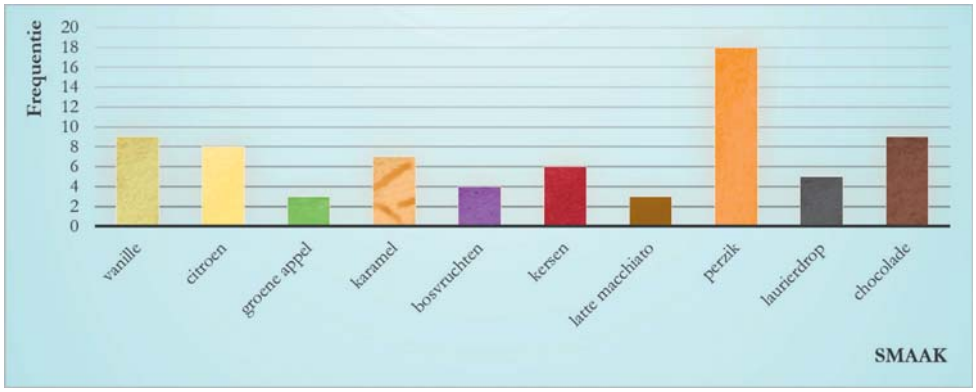


IV Staafdiagram

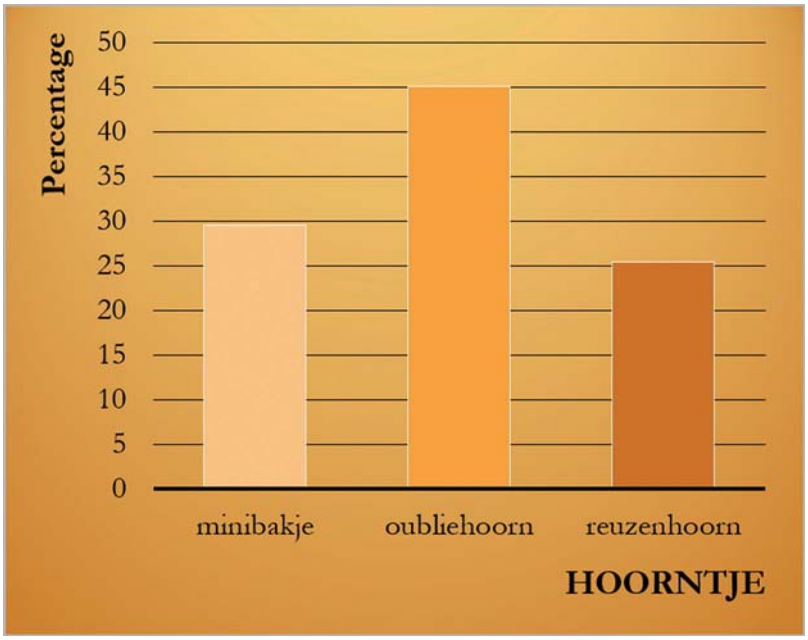
Nominaal Ordinaal Interval Ratio

Figuur 1.3 bestaat uit staven die het aantal ijsjes in een smaakcategorie aangeven – of de hoogte van het percentage, afhankelijk van onze voorkeur. Een staafdiagram doet dus hetzelfde als een cirkeldiagram, maar dan net even anders. Heb je een ordinale variabele, dan kun je de laagste rang links zetten en de hoogste rechts (figuur 1.4). Vandaar dat ik in zo'n geval een staafdiagram aanraad.

FIGUUR 1.3 Staafdiagram van de smaakvariabele



FIGUUR 1.4 Staafdiagram van de hoorntjesvariabele



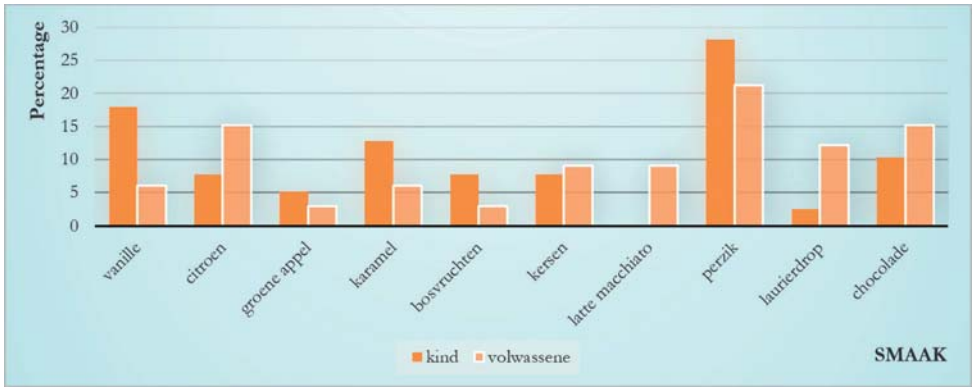
Samengesteld staafdiagram

Kijk je liever apart naar kinderen en volwassenen? Dat kan. Figuur 1.5 is een *samengesteld staafdiagram*, waarin alle smaken zijn uitgesplitst naar de leeftijd van de klant. Hier en daar komen wat verschillen bovendrijven. Simpele zoete smaken als karamel en vanille waren meer in trek bij kinderen, terwijl vooral volwassenen afkwamen op complexere smaken als latte macchiato en de kruidige laurierdrop. Fruit was doorgaans in beide groepen geliefd. Let wel: we praten hier slechts over een steekproef van 72 klanten, dus de volgende dag kan het beeld weer anders zijn.

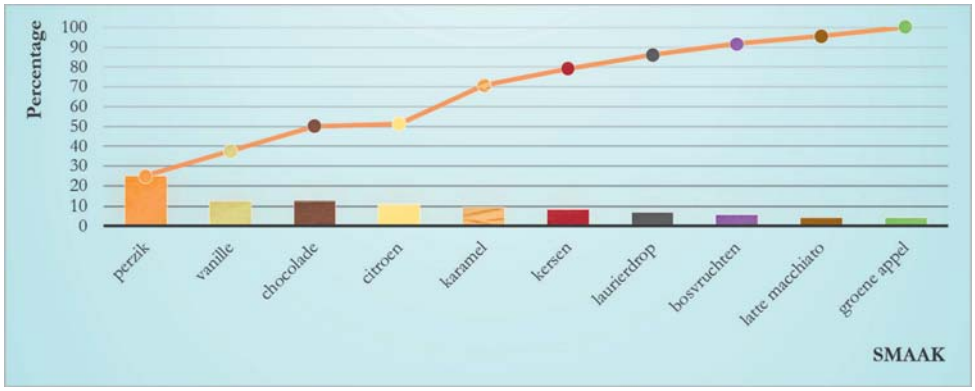
Paretodiagram

Voor zijn *nominale* variabele had Peach's Castle ook nog de optie om een *paretodiagram* te maken (figuur 1.6). In dit speciale staafdiagram zijn de categorieën geordend naar frequentie, van hoog naar laag. Perzikparade kwam links te staan als populairste smaak, gevolgd door vanille voor dummies en truffelchocolade. Enzovoort. We maken dus de smaakvariabele dus 'zogenaamd ordinaal' ... en de lijn in het diagram geeft de cumulatieve percentages aan. Wat blijkt daaruit: de drie populairste smaken waren al verantwoordelijk voor 50% van het verkochte ijs. Daarvan kan het bedrijf dus maar beter een flinke voorraad aanleggen.

FIGUUR 1.5 Samengesteld staafdiagram van de smaakvariabele



FIGUUR 1.6 Paretodiagram van de smaakvariabele

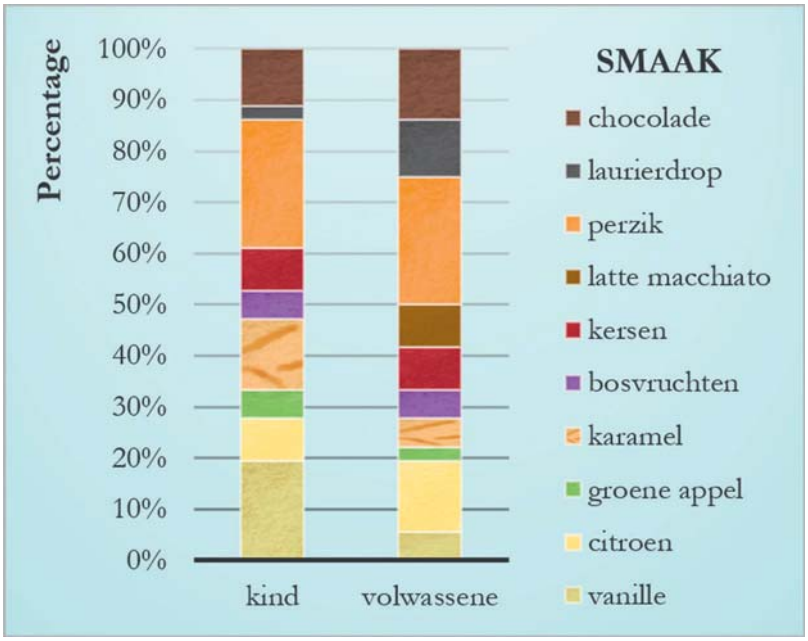


V Stapeldiagram

Nominaal Ordinaal Interval Ratio

Nog een variant op het standaard staafdiagram. Bestudeer je één groep, dan heb je maar één staaf, waarvan de stukjes steeds één categorie aangeven (zoals een ijssmaak). Hoe groter de categorie, hoe groter het desbetreffende stukje. Stapeldiagrammen brengen dus eigenlijk relatieve aantallen in beeld. Met een samengesteld stapeldiagram kun je weer meerdere groepen vergelijken, zoals kinderen en volwassenen. Opnieuw zien we dan natuurlijk dat kinderen vaker karamel en vanille hebben gekozen dan volwassenen, die op hun beurt weer vaker voor laurierdrop en latte macchiato zijn gegaan. Je mag trouwens ook staven maken van de absolute frequenties in plaats van de percentages. Dit kan handig zijn als je niet zozeer de groepen met elkaar wilt vergelijken, maar juist wilt laten zien hoeveel volwassenen en kinderen er zijn in totaal.

FIGUUR 1.7 Samengesteld stapeldiagram van de smaakvariabele



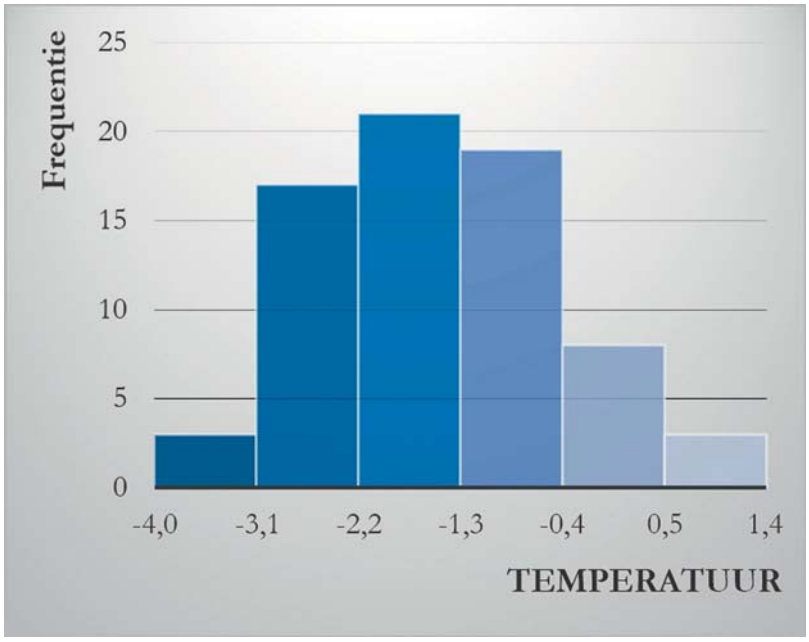
VI Histogram

Nominaal Ordinaal **Interval** **Ratio**

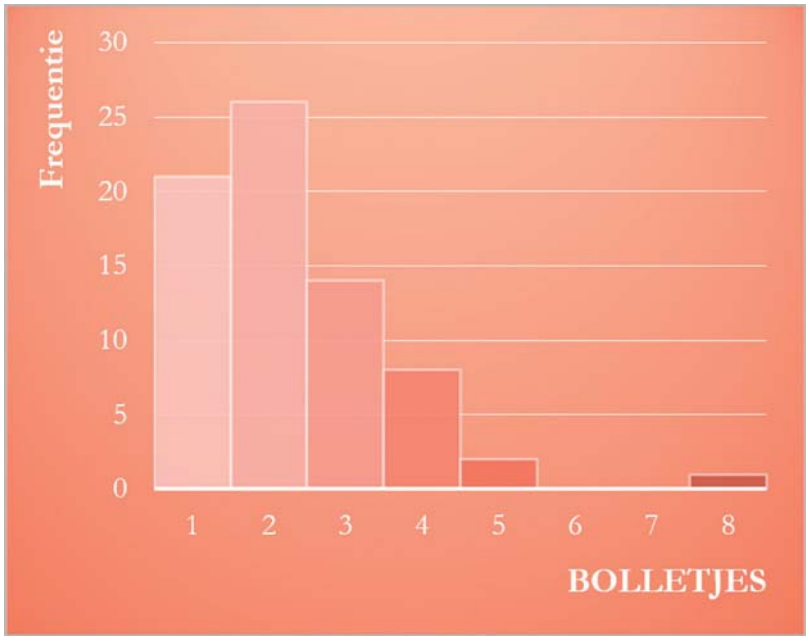
Hiermee komen we aan bij de belangrijkste grafiek voor kwantitatieve data. Het histogram is weer een speciaal staafdiagram. Het volgt hetzelfde principe als een stamdiagram: waarden die op elkaar lijken, worden samengenomen in één klasse (staaf). Figuur 1.8 laat dit zien voor de temperatuurvariabele. De eerste balk van links telt alle ijsjes met temperaturen tussen de –4,0 en –3,1 graden. Dat zijn er 3 (van –4,0, –3,3 en –3,2 graden, zoals we eerder in tabel 1.5 hebben gezien). De tweede balk telt het aantal ijsjes met temperaturen tussen de –3,1 en –2,2 graden en ga zo maar door. Elke klasse heeft dezelfde breedte (hier 0,9 graden). We krijgen zo een goed beeld van de tendensen in de verdeling.

Belangrijk is dat we de klassen niet te smal nemen, want dan kunnen we nauwelijks meer twee scores samen groeperen in dezelfde balk (figuur 1.10). Dat schiet niet op. Ook mogen de klassen niet te breed zijn: als we de eerste klasse definiëren als '–4,0 tot en met 1,0 graden', krijgen we één groot blok waar alle 71 ijsjes in zitten (figuur 1.11). Daar hebben we al helemáál niets aan. 😐 Het eerste histogram is prima: diverse onderscheidende balkjes staan mooi tegen elkaar aan.

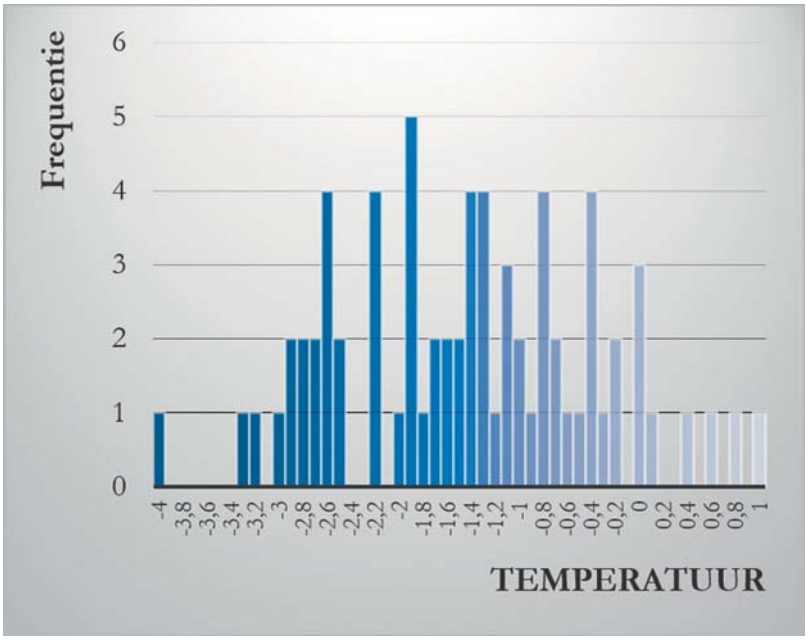
FIGUUR 1.8 Histogram van de temperatuurvariabele



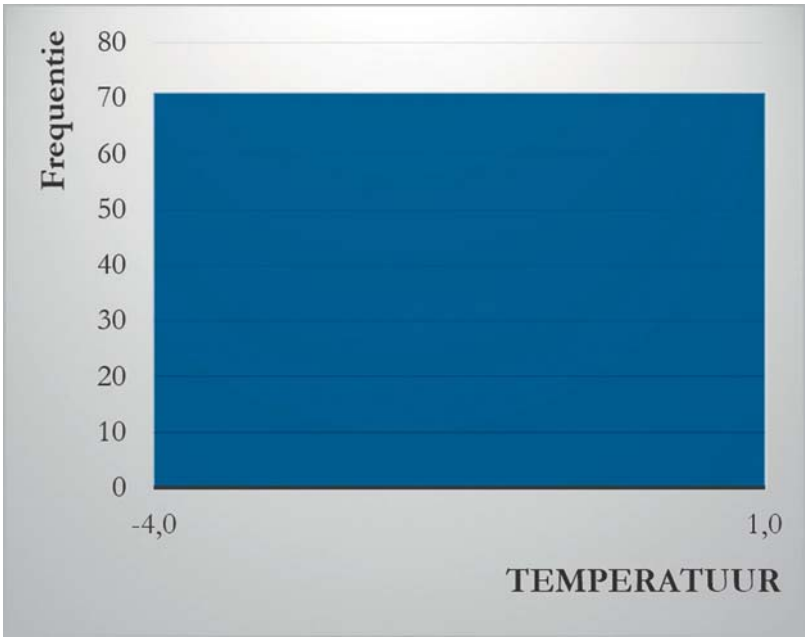
FIGUUR 1.9 Histogram van de bolletjesvariabele



FIGUUR 1.10 Te smalle klassen



FIGUUR 1.11 Te brede klassen

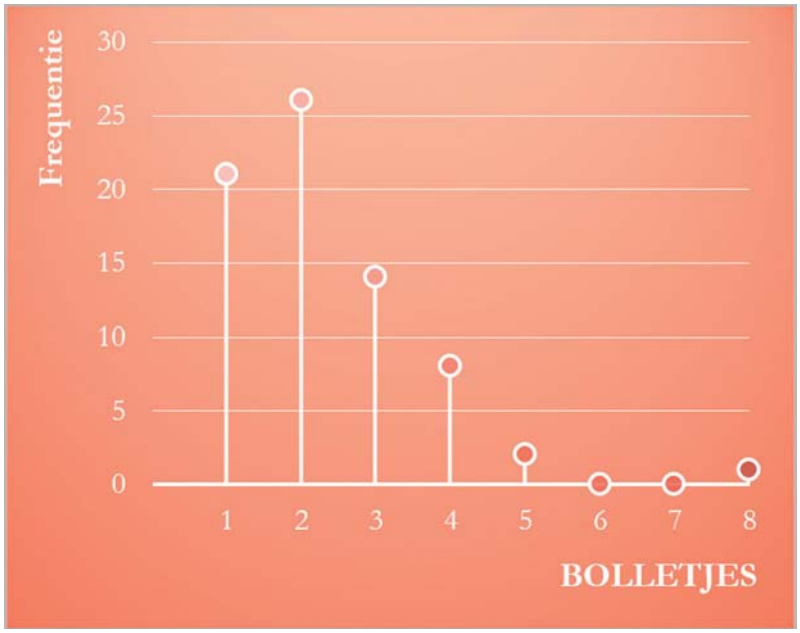


VII Naalddiagram

Nominaal Ordinaal **Interval** **Ratio**

Eigenlijk precies hetzelfde als een histogram, maar dan voor zogenoemde *discrete* variabelen. Dat zijn bijvoorbeeld variabelen met waarden die alleen uit gehele getallen bestaan, zoals de bolletjesvariabele (je kunt als klant geen halve bol ijs bestellen). We gaan hier in hoofdstuk 5 verder op in. Figuur 1.12 doet hetzelfde als figuur 1.9. De staven zijn echter naaldjes, om aan te geven dat er niets tussen de gehele getallen zit. Af en toe is dit zeker handig om te benadrukken. In de rest van het boek zal ik voornamelijk werken met gewone histogrammen.

FIGUUR 1.12 Naalddiagram van de bolletjesvariabele

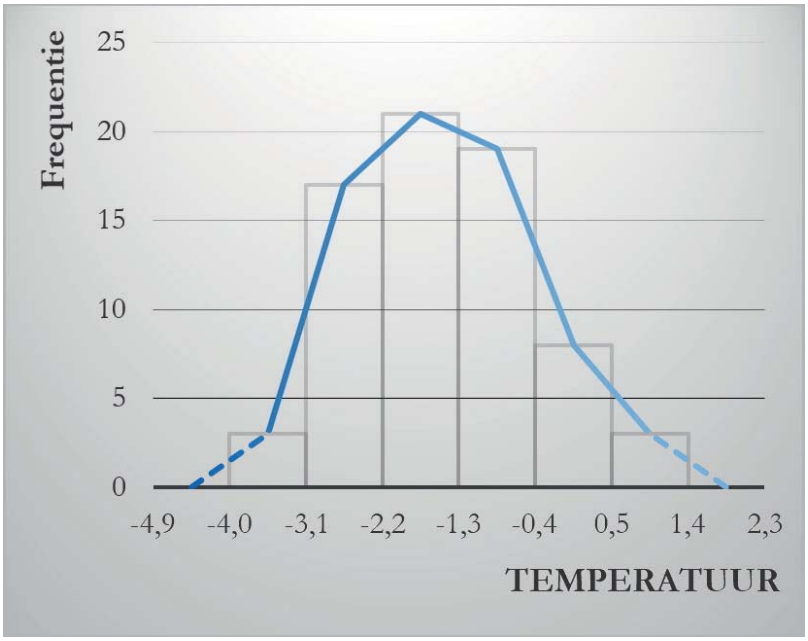


VIII Frequentiepolygoon

Nominaal Ordinaal **Interval** **Ratio**

De tegenpool van het naalddiagram. Frequentiepolygonen zijn juist geschikt voor continue variabelen, die waarden met oneindig veel decimalen kunnen hebben. Neem nog een keer de temperatuurvariabele. Om een frequentiepolygoon te maken, pak je het histogram en je verbindt de toppen van de staven met elkaar (je koppelt het midden van elke top, dus het midden van de klasse). Tot slot doe je net alsof er links én rechts nog een extra staaf zit met een frequentie van 0 (dus ter hoogte van de x-as) en je trekt de verbindingslijn gestippeld door naar die extra staven. Nu haal je het histogram weg *et voilà*: de lijn die overblijft is je polygoon.

FIGUUR 1.13 Frequentiepolygoon (het histogram laten we normaal weg)



Cumulatieve frequentie-polygoon

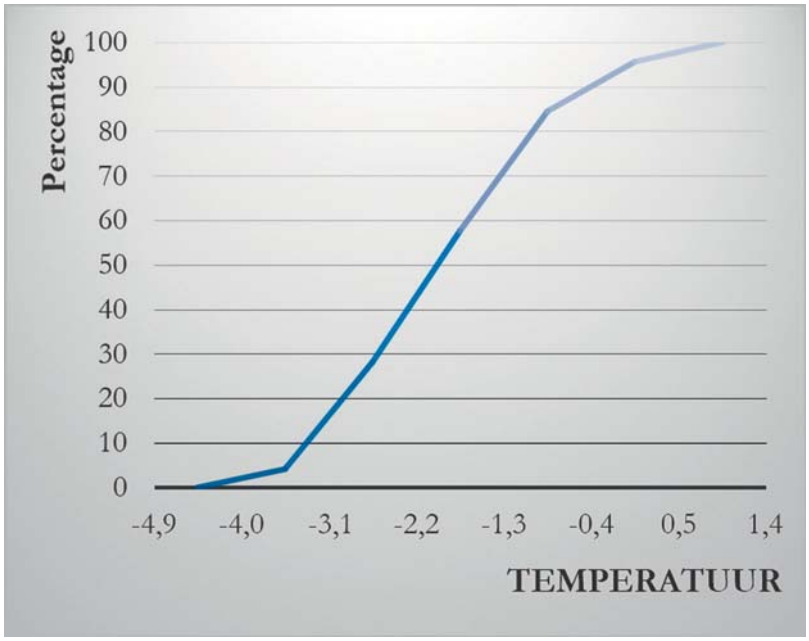
Het kan ook interessant zijn om een *cumulatieve frequentiepolygoon* te maken. Daarvoor moet je na elke staaf in het histogram vaststellen hoeveel procent van de scores je al hebt gehad:

TABEL 1.8 Frequentieverdeling van de temperatuurvariabele, met dezelfde klassenbreedten als figuur 1.8

		Frequentie	Percentage (reguliere polygoon)	Cumulatief percentage (cumulatieve polygoon)
Temperatuur	tussen −4,0 en −3,1	3	4,2	4,2
	tussen −3,1 en −2,2	17	23,9	28,2
	tussen −2,2 en −1,3	21	29,6	57,7
	tussen −1,3 en −0,4	19	26,8	84,5
	tussen −0,4 en 0,5	8	11,3	95,8
	tussen 0,5 en 1,4	3	4,2	100,0
	Totaal	**71**	**100,0**	

Figuur 1.14 beeldt die cumulatieve percentages af in een lijn. Dat is de cumulatieve polygoon.

FIGUUR 1.14 Cumulatieve frequentiepolygoon



IX Boxplot

Nominaal Ordinaal **Interval** **Ratio**

Een boxplot laat zien hoe een kwantitatieve variabele verdeeld is op basis van kwartielen. Maar dan moet ik natuurlijk eerst uitleggen wat kwartielen zijn. Dat gaat gemakkelijker in de volgende paragraaf, dus we komen nog terug op het boxplot.

Zijn er nog vragen?

De percentages in tabel 1.2 tellen niet mooi op tot 100%. Je hebt zeker een klunzige typefout gemaakt?
Nee hoor, dat komt gewoon door afronding.

Vincenzo, ik heb een computerprogramma gebruikt om een frequentietabel te maken. Mijn tabel bevat ook een kolom met het opschrift Valid Percent. *Wat is dat?*
Als er van een persoon geen geldige score in het databestand staat, spreken we van een ontbrekende waarde *(missing value)*. In de berekening van het *Valid Percent* zijn alleen de personen mét een geldige score meegeteld.

Ik wil een histogram maken, maar wat moet ik doen met scores die precies op de grens van een staaf zitten? Stel, ik heb een ijsje van −2,2 graden: moet ik dat dan nog meetellen in de balk 'tussen −3,1 en −2,2' of juist in de volgende balk 'tussen −2,2 en −1,3'?
Dat maakt niet uit, mits je consequent bent. In figuur 1.8 is die −2,2 nog meegeteld in de eerste balk (het computerprogramma Excel doet dat standaard; zie paragraaf 1.6). Heb je nog zo'n grensgeval, bijvoorbeeld −0,4 graden, dan tel je die dus op dezelfde manier: niet bij de balk 'tussen −0,4 en 0,5', maar bij de balk 'tussen −1,3 en −0,4'.

Nog even over dat histogram. Hoe kan ik gemakkelijk een klassenbreedte kiezen die geschikt is voor mijn gegevens? Heb je daar een tip voor?
Simpele computerprogramma's kijken vaak naar het totale aantal scores, trekken hier de wortel uit en ronden de uitkomst naar boven af. Bij de 72 klanten in het voorbeeld van Peach's Castle wordt dat $\sqrt{72} = 8{,}49 \rightarrow 9$ klassen, dus 9 staven. Vervolgens moet je dan kijken wat het bereik van je variabele is (dus: wat is het verschil tussen de hoogste en de laagste score?) en dit in 9 gelijke stukjes hakken (deel het bereik door 9).
Over het algemeen werkt dit aardig. Heeft je variabele een klein bereik, dan krijg je echter al snel overdreven veel staven. Heeft hij juist een groot bereik, dan worden uitschieters soms onzichtbaar gemaakt (zie paragraaf 1.4). De betere statistieksoftware houdt gelukkig ook rekening met het bereik. ☺ Excel en SPSS (zie paragraaf 1.6) doen dat tegenwoordig allebei.

1.3 Wie houdt er niet van vanille? Centrummaten

Het is één ding om een verzameling scores overzichtelijk in beeld te krijgen, maar het zou ook erg nuttig zijn om die verzameling samen te vatten met een paar kengetallen. Om het rekenwerk te vereenvoudigen, beperken we ons in deze paragraaf tot de eerste tien klanten. Hier komen hun scores nog een keer, plus een paar handige tabellen:

TABEL 1.9 Het ingekorte gegevensbestand voor deze paragraaf

Nummer ijsje	Smaak	Hoorntje	Temperatuur	Bolletjes
1	perzikparade	reuzenhoorn	−0,4	3
2	vanille voor dummies	minibakje	−2,2	1
3	vanille voor dummies	???	???	4
4	perzikparade	oubliehoorn	0,4	2
5	perzikparade	oubliehoorn	−1,9	2
6	perzikparade	minibakje	−1,4	2
7	truffelchocolade	oubliehoorn	−2,8	3
8	perzikparade	oubliehoorn	−2,5	2
9	perzikparade	reuzenhoorn	−0,3	8
10	vanille voor dummies	oubliehoorn	−1,4	3

TABEL 1.10 Frequentietabel van de smaakvariabele

		Frequentie	Percentage
Smaak	truffelchocolade	1	10,0
	perzikparade	6	60,0
	vanille voor dummies	3	30,0
	Totaal	10	100,0

TABEL 1.11 Frequentietabel van de hoorntjesvariabele

		Frequentie	Percentage	Cumulatief percentage
Hoorntje	minibakje	2	22,2	22,2
	oubliehoorn	5	55,6	77,8
	reuzenhoorn	2	22,2	100,0
	Totaal	9	100,0	

TABEL 1.12 Stamdiagram van de temperatuurvariabele

Temperatuur	
−2,	2 5 8
−1,	4 4 9
−0,	3 4
0,	4

TABEL 1.13 Stamdiagram van de bolletjesvariabele

Bolletjes	
1,	0
2,	0 0 0 0
3,	0 0 0
4,	0
8,	0

Om een variabele samen te vatten, bepalen we om te beginnen vaak een *centrummaat*. Wat dat is, wordt waarschijnlijk het snelst duidelijk als we drie populaire centrummaten op een rijtje zetten:

I Modus (*Mo*)
Dit is de score die *het vaakst voorkomt*. De modi van onze variabelen wijzen zich vanzelf:

- Smaak: de modus is ‘perzikparade’. Oftewel, ook in deze kleinere groep gingen de meeste klanten voor de specialiteit van Peach’s Castle.
- Hoorntje: de modus is ‘oubliehoorn’. De meeste klanten kozen dus een hoorn van gematigde grootte.
- Temperatuur: er is in dit geval één score die vaker dan één keer voorkomt, namelijk –1,4 graden. De meest voorkomende temperatuur waarop de ijsjes geserveerd werden, lag derhalve iets onder het vriespunt (0). Zo hoort het natuurlijk ook.
- Bolletjes: de modus is 2. Anders gezegd, de meeste klanten kochten twee bolletjes ijs.

Je ziet dat de modus (meervoud: modi) op *elk meetniveau* gerapporteerd kan worden. Tenslotte zegt hij alleen iets over de *frequentie* van de scores.

II Mediaan (*M*)

Dit is de waarde die *in het midden ligt als we de scores ordenen van laag naar hoog*. 50% van de scores ligt dan onder de mediaan en 50% erboven. Laten we eens kijken:

- Smaak: lukt het om hiervan een mediaan te rapporteren? Hoe sorteren we de scores dan? We kunnen net zo goed de vanille-ijsjes bovenaan zetten in plaats van de chocolade-ijsjes. Bij een nominale variabele betekent een mediaan dus niets; hij is willekeurig, afhankelijk van de betekenisloze volgorde waarin we de waarden van de variabele zetten.
- Hoorntje: de frequentietabel heeft de scores al voor ons gesorteerd. Nu moeten we nog de middelste vinden. Hier is de volledige lijst:

TABEL 1.14 Hoorntjes van de 9 klanten, geordend

Hoorntje	Rang
minibakje	1e
minibakje	2e
oubliehoorn	3e
oubliehoorn	4e
oubliehoorn	5e
oubliehoorn	6e
oubliehoorn	7e
reuzenhoorn	8e
reuzenhoorn	9e

Met name als de lijst langer wordt, is het niet meer zo simpel om hem helemaal uit te schrijven en de middelste rij aan te wijzen. Daarom het volgende trucje: neem het totale aantal scores (N) plus 1, en deel dit door 2 om het midden te vinden. $\frac{9+1}{2} = 5$. Let erop dat dit de *rang* van de middelste score is, niet de waarde! Welk hoorntje koos de 5e klant? Dat blijkt een oubliehoorn te zijn. We kunnen dit tellen in de frequentietabel: de 5e klant zit in de categorie 'oubliehoorn'. Ook zien we dat bij deze categorie het cumulatieve percentage over de 50% heen gaat. Het ijsje dat precies op de grens van 50% zit, moet zich dus in deze categorie bevinden.

- Temperatuur: gebruik het stamdiagram om de middelste score aan te wijzen (pas wel op dat de laagste scores bij getallen onder nul niet voorraan staan!). Het is wederom de 5e, met een waarde van –1,4. Hetzelfde als de modus dus.
- Bolletjes: voor deze variabele hebben we niet 9 maar 10 scores (dit komt doordat één klant moeilijk deed bij de meting van zijn hoorntje en

temperatuur, weet je nog?). De trucformule zegt: *rang* = $\frac{10+1}{2}$ = 5,5. Verdikkeme, de 'middelste' persoon zit tussen de vijfde en de zesde in! Voor de mediaan moeten we dan maar het gemiddelde van deze scores nemen. Het gemiddelde van 2 en 3 bolletjes is $M = \frac{2+3}{2}$ = 2,5.

De mediaan vertelt ons dus iets over de *frequentie* en de *ordening* van de scores. We kunnen deze rapporteren bij variabelen die *minstens ordinaal* zijn.

III Rekenkundig gemiddelde ($\overline{X}$)

Dit is het *zwaartepunt* van de scoreverdeling. Om het rekenkundig gemiddelde te berekenen, tellen we alle losse scores (uitgedrukt als X) bij elkaar op en delen we ze door het totale aantal scores (uitgedrukt als N). In een formule ziet de berekening er zo uit: $\overline{X} = \frac{\sum X_i}{N}$. Deze notatie zal ik even uitleggen:

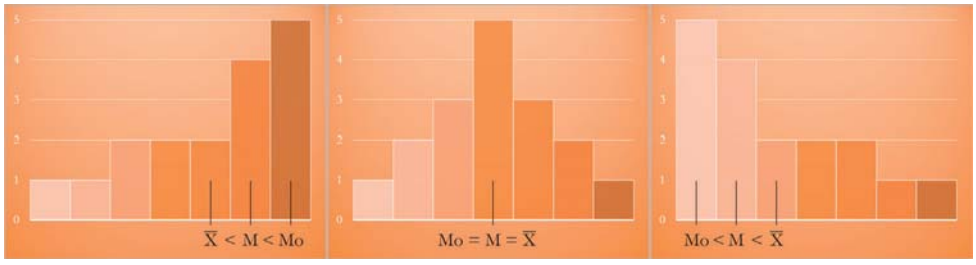
- Elke X hoort bij een specifieke klant, aangeduid met de letter i. Dit is gewoon een nummer. X_1 is de score van klant nummer 1, X_2 is de score van klant nummer 2 enzovoort.
- De $\sum$ (sigma) is een Griekse 'S' en staat voor 'som', dus hij draagt ons in dit geval op om alle X'en die we hebben te sommeren.

Rekenen maar:

- Temperatuur: alle temperaturen tellen samen op tot $\sum X_i = -12{,}5$ (controleer het gerust). Dus wordt het gemiddelde $\overline{X} = \frac{-12{,}5}{9} = -1{,}39\ldots$ bijna hetzelfde als de modus en mediaan.
- Bolletjes: de som van alle klanten is $\sum X_i = 30$. Dat maakt $\overline{X} = \frac{30}{10} = 3$. Kortom, gemiddeld kopen de klanten 3 bolletjes ijs.

Het gemiddelde bevat informatie over de *frequentie*, de *ordening* en de *waarde* van de scores; we rekenen immers met die waarden. Daarom kunnen we het gemiddelde rapporteren als een variabele *minstens van intervalniveau* is. Bij een categorische variabele betekent het gemiddelde niets: 'Wat was de gemiddelde smaak?' Juist, ja.

FIGUUR 1.15 Verdelingsvormen: scheef naar links, symmetrisch, scheef naar rechts



Bij de bolletjesvariabele blijken het gemiddelde (3), de mediaan (2,5) en de modus (2) nogal te verschillen – terwijl ze voor de temperatuur bijna identiek zijn (ongeveer –1,4). Hoe kan dat? Het komt doordat de verdeling van

de temperatuur redelijk *symmetrisch* is; hoge en lage scores houden elkaar in balans, zodat onze centrummaten op elkaar lijken. Echter, de verdeling van de bolletjes is *scheef* (maak maar eens een histogram zoals figuur 1.9; het ziet er bijna hetzelfde uit). Figuur 1.15 laat je zien wat er in het algemeen gebeurt. Scheefheid verwijst naar de *staart* van de scoreverdeling, dus de linkergrafiek is scheef naar links en de rechter is (net zoals de bolletjesvariabele) scheef naar rechts. In het geval van linkse scheefheid zijn er wat lage scores die soms de mediaan iets omlaag trekken, en het gemiddelde vaak (maar niet altijd) nog iets meer. Rechtse scheefheid houdt in dat we wat hoge scores hebben die de mediaan soms iets omhoogtrekken, en die met name het gemiddelde kunnen vertekenen. De mediaan is dan meestal een redelijk compromis. We kunnen bij scheve verdelingen dus de voorkeur geven aan de mediaan, zelfs als onze variabele kwantitatief is.

Scheve verdeling

Uitschieter

Je zou zelfs kunnen zeggen dat de verdeling van de bolletjes een *uitschieter* bevat: een extreem lage of hoge score die niet lijkt op de rest. Er is één enkele klant die maar liefst 8 bolletjes heeft besteld, veel meer dan gebruikelijk. Als we het gemiddelde berekenen *zonder* deze klant, krijgen we $\overline{X} = 2{,}44$. Dat lijkt veel meer op de huidige modus en mediaan en geeft een beter beeld van het aantal bolletjes dat klanten in het algemeen kopen. Ook uitschieters vertekenen het gemiddelde dus soms.

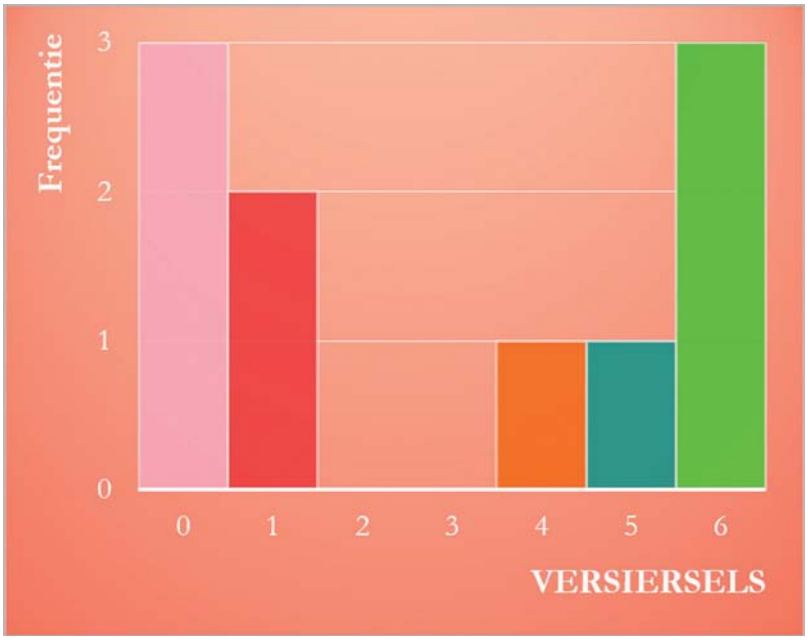
Iets om op te letten

Tweetoppige of bimodale verdelingen

Sommige verdelingen zijn *tweetoppig*; we noemen ze ook wel *bimodaal*, omdat we twee *(bi)* modi nodig hebben om de verdeling goed te beschrijven – of twee medianen en twee gemiddelden. Een voorbeeldje (figuur 1.16): de hoeveelheid versieringen die klanten graag op hun ijsje willen, is misschien wel tweetoppig verdeeld. Volwassenen nemen niets of hooguit wat gehakte nootjes, maar kinderen willen het liefst *alles*: aardbeiensaus, chocoladeganache, discodip, een parasolletje … We hebben dus twee verschillende *subgroepen*.

Subgroepen

FIGUUR 1.16 Tweetoppige verdeling



Zijn er nog vragen?

Je drukte je nogal voorzichtig uit toen je de modus, de mediaan en het gemiddelde vergeleek bij scheve verdelingen. Geldt wat we in figuur 1.15 zien dan niet altijd?
Dat klopt – er zijn uitzonderingen. Sommige rechtsscheve verdelingen hebben zelfs een gemiddelde dat *kleiner* is dan de modus. Dat kan gebeuren als de staart heel weinig scores bevat en het gebied links van de top juist heel veel.

Andersom heeft een linksscheve verdeling soms een gemiddelde dat *groter* is dan de modus.

Ik heb ergens iets gelezen over uitbijters. Zijn dat gewoon uitschieters?
Ja, een uitschieter wordt soms ook een *uitbijter* genoemd. Persoonlijk vind ik dat maar een rare term. Waar heeft-ie dan in gebeten? 😐

Uitbijter

1.4 Wie heeft er perzik besteld? Spreidingsmaten

Als je een kwantitatieve variabele hebt, kun je ook in kaart brengen hoe groot de *spreiding* is: bestellen alle klanten vrijwel evenveel bolletjes ijs, of verschillen ze hierin behoorlijk? Deze informatie kan minstens even belangrijk zijn voor Peach's Castle, aangezien de ijssalon altijd genoeg op voorraad moet hebben.

Spreiding

Een manier om spreiding te benaderen, is de vraag hoeveel de scores verschillen van het gemiddelde. Laten we de bolletjesvariabele bekijken. De getallen vind je (gesorteerd) in tabel 1.13. We berekenen nu van elke score de afwijking van het gemiddelde (3), en vervolgens tellen we al die afwijkingen bij elkaar op. Oftewel:

$$\sum(X_i - \overline{X}) =$$
$$(1-3) + (2-3) + (2-3) + (2-3) + (2-3) +$$
$$(3-3) + (3-3) + (3-3) + (4-3) + (8-3) =$$
$$\mathbf{0}$$

O, drommels. Dat hadden we kunnen weten. Tenslotte zit het gemiddelde precies tussen alle scores in: het ene verschil met het gemiddelde is negatief, het andere positief, en zo heffen ze elkaar perfect op.
De oplossing? We *kwadrateren* alle verschillen voordat we ze optellen. Daar gaan we weer:

$$\sum(X_i - \overline{X})^2 =$$
$$(1-3)^2 + (2-3)^2 + (2-3)^2 + (2-3)^2 + (2-3)^2 +$$
$$(3-3)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2 + (8-3)^2 =$$
$$\mathbf{34}$$

Even vooruitlopen op hoofdstuk 3: dit noemen we de *kwadratensom* of ook wel de *variatie* in het aantal bolletjes ijs. Ze is een beetje abstract, maar ze vertelt je hoeveel de klanten samen variëren van het gemiddelde. Als iedereen erg op elkaar lijkt is de variatie klein.

Kwadratensom
Variatie

Wat is nu de *gemiddelde (gekwadradeerde) afwijking* per klant? Om hierachter te komen, delen we de kwadratensom door het aantal klanten (N) minus 1. Deze gemiddelde afwijking staat bekend als de *variantie*.

Variantie

$$s_X{}^2 = \frac{\Sigma(X_i - \overline{X})^2}{N - 1} = \frac{34}{10 - 1} = 3{,}78$$

Standaard-afwijking

Standaard-deviatie

De variantie geeft aan hoeveel bolletjes de gemiddelde klant afwijkt van het gemiddelde – op een gekwadrateerde schaal. Dit is nog steeds lastig te interpreteren; ik bedoel, wat is een gekwadrateerd bolletje? Kun je dat nog wel eten? Laten we dus terugkeren naar onze oorspronkelijke schaal. Als we ten slotte de *wortel* nemen uit deze variantie, krijgen we de zogenoemde *standaardafwijking* of *standaarddeviatie* (betekent precies hetzelfde):

$$s_X = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{N - 1}} = \sqrt{3{,}78} = 1{,}94$$

Oftewel: gemiddeld wijkt het aantal bolletjes ijs dat de klanten bestelden met 1,94 bolletjes af van het gemiddelde. De definitie van de standaarddeviatie is dan ook de gemiddelde afwijking van het gemiddelde.

Een nadeel van de standaardafwijking is dat deze net zoals het gemiddelde vertekend kan worden door scheve verdelingen en uitschieters. Die ene klant die 8 bolletjes kocht, zat ver boven het gemiddelde; daardoor blaast hij de standaardafwijking flink op – de gemiddelde afwijking van het gemiddelde. We zagen eerder al dat de bolletjesvariabele onze voorkeur doet verschuiven naar de mediaan als centrummaat. In dat geval is het gepast om ook een andere spreidingsmaat te kiezen: de *interkwartielafstand*, vaak afgekort tot IQR (van het Engelse *interquartile range*).

Interkwartiel-afstand

Als we de scores ordenen van hoog naar laag, krijgen we de lijst in tabel 1.15. We kunnen nu de dataset opsplitsen in vier 'blokken' die elk 25% van de scores bevatten. Elke tussengrens noemen we een *kwartiel*. (Sommige statistici bedoelen met kwartielen de blokken zelf, maar mijn termen schijnen het meest gebruikelijk te zijn.)

Kwartiel

- Het *eerste kwartiel* (Q_1) is de mediaan van de 50% laagste scores. 25% van alle scores ligt eronder en 75% erboven.
- Het *tweede kwartiel* (Q_2) is weer de 'gewone' mediaan.
- Het *derde kwartiel* (Q_3) stelt de mediaan van de 50% hoogste scores voor. 75% van alle scores ligt eronder en 25% erboven.

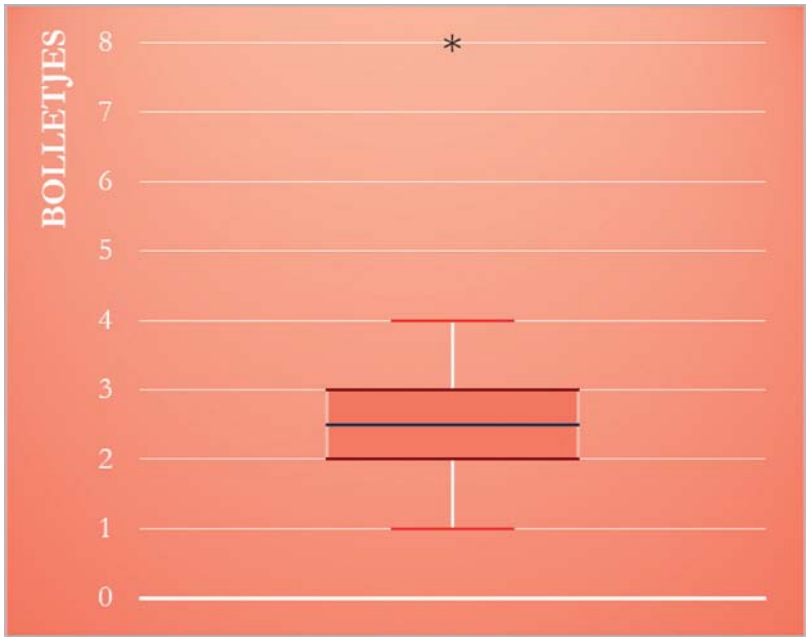
TABEL 1.15 Kwartielen

8	
4	
3	Q_3
3	
3	Q_2: mediaan (2,5)
2	
2	
2	Q_1
2	
1	

Het *boxplot* in figuur 1.17 visualiseert de kwartielen. Q_1 en Q_3 vormen de onder- en bovenrand van de box, en Q_2 is de lijn die erbinnen doorheen loopt. Ten slotte hebben we de zogenaamde 'snorharen' in het roze: zij beschrijven de laagste score (minimum) en de hoogste (maximum). Blijkbaar wordt de klant met 8 bolletjes geteld als een uitschieter: die staat aangeduid met een ster die buiten de box valt. We nemen daarom de eerstvolgende (4) als de snorhaar voor het maximum.

Boxplot

FIGUUR 1.17 Boxplot



Je kunt zeggen dat het boxplot een visuele samenvatting biedt van de data. Deze *vijfgetallensamenvatting* kan ook in een lijst worden gezet (minimum, Q_1, mediaan, Q_3, maximum):

Vijfgetallen-samenvatting

1 2 2,5 3 4

Goed. Wat is nu die IQR waar ik het zojuist over had? Merk op dat we tussen Q_1 en Q_3 twee keer 25% van de scores vinden – oftewel de middelste 50%. Met andere woorden, het verschil tussen Q_1 en Q_3 bevat de helft van de scores. Dit verschil noemen we de interkwartielafstand.

$$IQR = Q_3 - Q_1 = 3 - 2 = 1$$

Kortom: de helft van de klanten ligt hooguit één bolletje uit elkaar. Als de IQR groot is, weten we dat de meest centrale scores al ver uiteenliggen. Merk op dat uitschieters geen invloed hebben op de IQR: die zou gelijk blijven al had de meest hongerige klant 15 bolletjes gekocht in plaats van 8.

Belangrijk detail: om de IQR te berekenen, trekken we twee kwantitatieve waarden van elkaar af. Dit slaat alleen ergens op als de variabele in kwestie kwantitatief is (wat dacht je immers van 'reuzenhoorn min oubliehoorn'?). Dat betekent dat we de IQR *niet* mogen rapporteren bij een ordinale variabele, ook al kunnen we daarvan wel de mediaan bepalen.
De laatste vraag is waarom ik die ene klant met zijn 8 bolletjes ijs beschouwde als een uitschieter. Daarvoor heb ik een gangbaar criterium gebruikt: het *1,5×IQR-criterium*. Het is vrij gemakkelijk. Eerst berekenen we twee grenzen (let op de subtiele verschillen in de twee formules):

1,5×IQR-criterium

$$Q_1 - (1{,}5 \times IQR) = 2 - (1{,}5 \times 1) = 2 - 1{,}5 = 0{,}5$$
$$Q_3 + (1{,}5 \times IQR) = 3 + (1{,}5 \times 1) = 3 + 1{,}5 = 4{,}5$$

De gedachte is dat je alle scores tussen 0,5 en 4,5 acceptabel mag noemen. Maar scores die *buiten dit bereik* vallen, zijn uitschieters! De klant die 8 bolletjes achteroversloeg, verdient daarom met verve het etiket van opwaartse uitschieter. Peach's Castle had vandaag geen klanten die minder dan een half bolletje ijs bestelden, dus we hebben geen neerwaartse uitschieter. (Dat zou ook tamelijk onrealistisch zijn; welke idioot bestelt er nu een kwart bolletje?)
Er bestaan meer manieren om potentiële uitschieters op te sporen, en het 1,5×IQR-criterium is slechts een stuk basisgereedschap om je op weg te helpen. Op termijn zul je wellicht je horizon verbreden, waarde lezer.

Extra: de variatiecoëfficiënt

Die standaardafwijking van 1,94 bolletjes (dus bijna 2), is dat nou veel of weinig? We zouden die eens kunnen vergelijken met het gemiddelde – en wel door de een door de ander te delen. De verhouding tussen de standaardafwijking en het gemiddelde staat bekend als de *variatiecoëfficiënt*.

$$c_v = \frac{s}{\overline{X}} = \frac{1{,}94}{3} = 0{,}65$$

Wat blijkt: vergeleken met het gemiddelde is de standaardafwijking 0,65 keer zo groot (dus iets meer dan half zo groot). Dat is aardig wat. Je moet het zo zien: de gemiddelde klant bestelt 3 bolletjes ijs, maar doorgaans wijken individuele klanten daar alweer een kleine 2 bolletjes van af. Stel je eens voor dat de gemiddelde klant 20 bolletjes bestelde (smakelijk!): dan zou een standaardafwijking van 2 daar schamel tegen afsteken. Relatief ten opzichte van het gemiddelde lijken de klanten in zo'n situatie nogal op elkaar. De variatiecoëfficiënt wordt (in dat absurdistische geval) gelijk aan $c_v = \frac{1{,}94}{20} = 0{,}097$, wat betekent dat de standaardafwijking slechts een kleine één tiende bedraagt van het gemiddelde.

Kortom, c_v is een leuke *relatieve* maat voor spreiding. Maar heb je er altijd iets aan? Nee: je kunt hem alleen gebruiken bij *rationiveau*. Tenslotte drukt c_v een verhouding uit. Bij intervalvariabelen slaat de variatiecoëfficient nergens op, want een gemiddelde dicht bij 0 hoeft dan niet klein te zijn; misschien loopt de schaal nog wel ver door onder 0. Denk maar weer aan de temperatuur van de ijsjes.

Zijn er nog vragen?

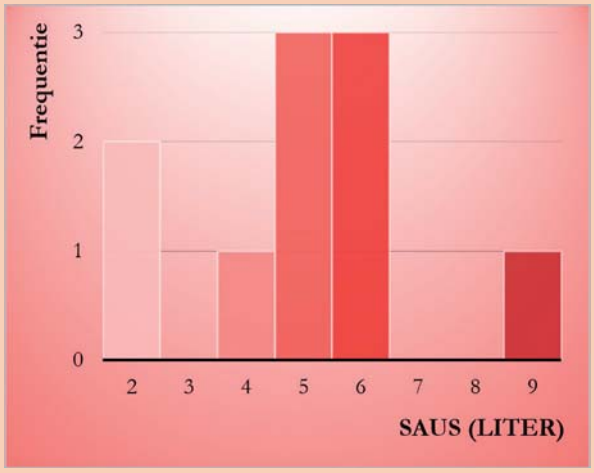
Bij de berekening van de variantie delen we de kwadratensom door N – 1. Waarom niet gewoon door N?
Als je het weten moet: geduld, Padawan. In hoofdstuk 5 leg ik het uit als extraatje. Voor alle andere lezers is het een priegelig detail dat je maar beter gewoon kunt aannemen.

Boxplot, bah. Onze taal verengelst alsmaar meer. Hebben we voor deze grafiek geen oer-Nederlandse naam?
Zeker wel. Hoe exotisch wil je het hebben? Sommigen noemen het een doosdiagram, anderen een kader-met-staafdiagram. Weer anderen – geen grap – noemen het een snorredoos …

Moet ik uitschieters altijd weglaten uit de vijfgetallensamenvatting?
Hier is geen vaste regel voor, dus je mag zelf kiezen. Mensen die deze lijstvorm gebruiken, tellen uitschieters in de verdeling soms wel als het minimum of maximum. Persoonlijk zou ik ze weglaten omdat ze de samenvatting vertekenen.

1.5 Kleurstof: lineaire transformaties

FIGUUR 1.18 Afzet van aardbeiensaus (ruwe variabele)



4	2
6	6
5	9
5	6
2	5

$\bar{X} = 5$
$s_X = 2{,}05$

Het állerlaatste stuk theorie van dit hoofdstuk (beloofd!) gaat over de vraag hoe we gegevens kunnen transformeren. Soms is het handig om de schaal van je variabele aan te passen. De term 'datatransformatie' klinkt intimiderend, maar – verrassing – je hebt het zelf al ontzettend vaak gedaan, beste lezer. Neem een emmer aardbeiensaus van 1 liter. Goh, hoeveel milliliter is dat eigenlijk? 1000 milliliter?
Gefeliciteerd: je hebt zojuist een datatransformatie uitgevoerd. Je veranderde het getal, maar de *betekenis* bleef hetzelfde. Het volume van die emmer is tenslotte niet gewijzigd. Andere typische datatransformaties zijn die van temperatuur (bijvoorbeeld van graden Celsius naar graden Fahrenheit) en van tijd (van seconden naar minuten, van minuten naar uren enzovoort).

1

Lineaire transformatie

Nu vragen we de bedrijfsleider van Peach's Castle hoeveel liter aardbeiensaus ze de afgelopen tien dagen hebben verkocht. De afzet per dag staat in figuur 1.18; het gemiddelde was 5 liter en de standaardafwijking 2,05. Transformaties komen voor in alle soorten en maten. Deze paragraaf bespreekt een paar eenvoudige vormen die bekendstaan als *lineaire transformaties*.

I Multipliceren

Als we de vorenstaande gegevens voorleggen aan een Amerikaanse lezer, zal die zachtjes mompelen: 'Hoeveel is een liter in vredesnaam?' Aan de overkant van de Atlantische Oceaan gebruikt men namelijk gallons. Voor Peach's Castle is dit vertaalprobleem een dingetje, want de ijssalon importeert zijn goddelijke aardbeiensaus uit de VS. Hoeveel moet men inkopen? Om daarachter te komen, moeten we onze schaal veranderen. 1 liter staat gelijk aan ongeveer 0,26417 *US liquid gallons*. Dus om de verkoop in gallons (G) uit te drukken, moeten we alle scores *vermenigvuldigen* met 0,26417:

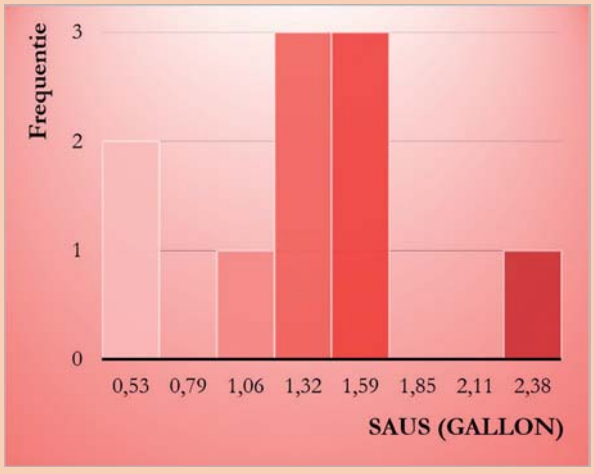
$$G_i = 0{,}26417 \times X_i$$

In het algemeen betekent multipliceren dus voorspelbaar dat we de oorspronkelijke (ook wel 'ruwe') variabele vermenigvuldigen met een bepaald getal dat we m kunnen noemen:

$$X'_i = m \times X_i$$

De i staat voor het individu waartoe de score behoort (in dit geval de dag); we voeren de berekening uit over alle individuele scores. Dus 0,26417 × 4 uit figuur 1.18, 0,26417 × 6, enzovoort.

FIGUUR 1.19 Afzet van aardbeiensaus (gemultipliceerd)



1,06	0,53
1,59	1,59
1,32	2,38
1,32	1,59
0,53	1,32

$\bar{G} = 1{,}32$
$s_G = 0{,}54$

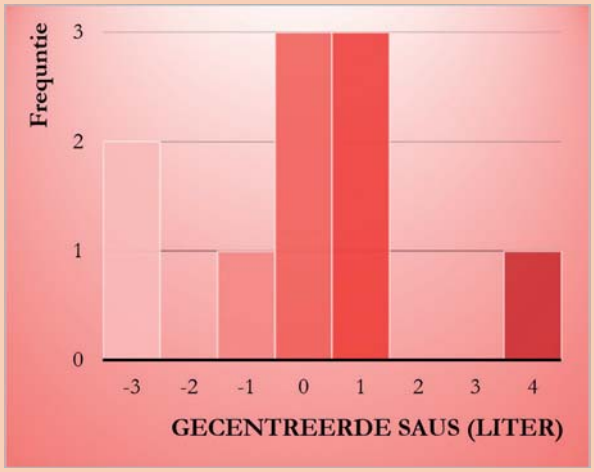
Wat er in dat geval gebeurt, zie je in figuur 1.19. Alle scores worden 0,26417 keer zo groot; hetzelfde geldt intuïtief voor het gemiddelde en de standaardafwijking. Als de gemiddelde verkoop 5 liter was, kun je immers gewoon zeggen dat dit gelijkstaat aan 0,26417 × 5 = 1,32085 gallons.
Dit helpt Peach's Castle te belissen hoeveel men voor de volgende maand moet bestellen op de website van de Amerikaanse producent. Kortom, met datatransformaties brengen we dezelfde informatie over op een andere manier.

II Centreren

Deze transformatie is een speciaal geval van een 'verschuiving', waarin we alle scores zo opschuiven dat het *gemiddelde* van de schaal 0 wordt. Dit is een eitje: we hoeven alleen het gemiddelde (hier 5) van alle scores af te trekken. De formule voor de nieuwe variabele is dus:

$$C_i = X_i - \overline{X}$$

FIGUUR 1.20 Afzet van aardbeiensaus (gecentreerd)



−1	−3
1	1
0	4
0	1
−3	0

$\overline{C} = 0$
$s_C = 2{,}05$

Zoals figuur 1.20 laat zien, gebeurt er niets met de vorm van de verdeling. De scores glijden simpelweg zijwaarts met 5 punten. Gecentreerde scores geven aan hoe ver scores afwijken van het gemiddelde; dit is wat hen informatief maakt. Neem bijvoorbeeld de eerste dag (–1): hierop lag de verkoop van aardbeiensaus 1 liter lager dan gemiddeld.

III Standaardiseren

Wie standaardiseert, zet de schaal om naar zogenoemde *z-scores*, die je nog veel zult gaan gebruiken. De formule is:

Z-score

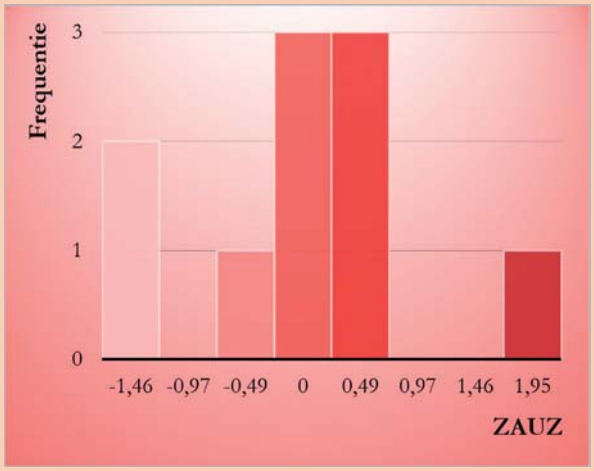
$$z_i = \frac{X_i - \overline{X}}{s_X}$$

Zie je dat er in de teller eerst wordt gecentreerd? Het *gemiddelde* van z-scores is daarom altijd 0. Vervolgens delen we de gecentreerde scores door s_X, wat eigenlijk een multiplicatie is. En wat doet deze multiplicatie met de standaardafwijking? Eerder, toen we de schaal vermenigvuldigden met 0,26417, zagen we de standaardafwijking 0,26417 keer zo groot worden. Dus als we de schaal nu *delen* door 2,055, wordt de standaarddeviatie:

$$s_z = \frac{2{,}055}{2{,}055} = 1$$

Precies: de *standaardafwijking* van z-scores is altijd 1. Aanschouw het in figuur 1.21.

FIGUUR 1.21 Afzet van aardbeiensaus (gestandaardiseerd)



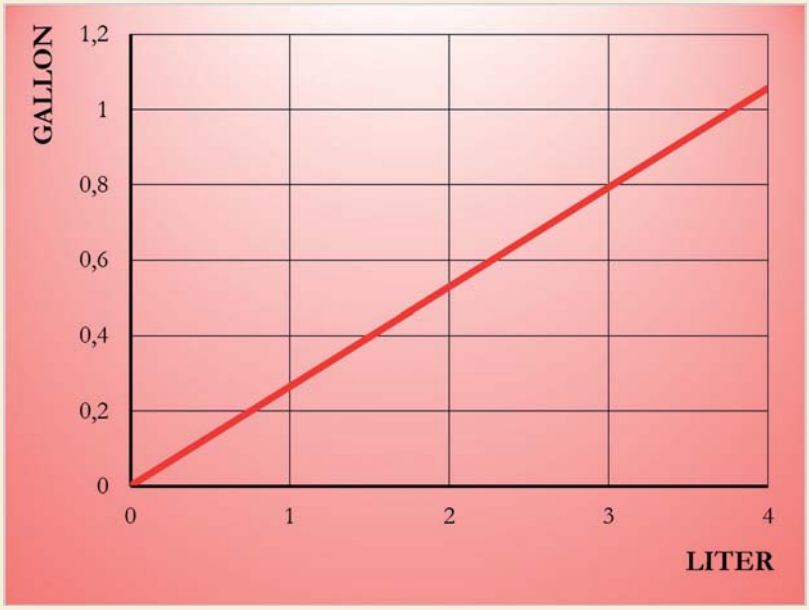
−0,49	−1,46
0,49	0,49
0	1,95
0	0,49
−1,46	0

$\bar{Z} = 0$
$s_Z = 1$

Z-scores geven aan hoeveel *standaardafwijkingen* een meting *boven of onder het gemiddelde* heeft gescoord. De achtste dag, waarop 9 liter saus werd verkocht, scoort 1,95: deze dag zit bijna twee standaardafwijkingen boven het gemiddelde (bijna twee keer 2,055), wat direct aangeeft dat er die dag veel meer aardbeiensaus over de toonbank ging dan gebruikelijk. Dit konden we in principe ook al zien op de ongetransformeerde schaal, alleen moesten we daarvoor het aantal verkochte liters op die dag vergelijken met het gemiddelde én de standaardafwijking. Een z-score vormt één handig getal.

Zijn er nog vragen?

Waarom noemen we dit soort transformaties eigenlijk lineair?
Als je een grafiek tekent, ziet het verband tussen de oude en de nieuwe variabele eruit als een rechte lijn. Bijvoorbeeld, van liters naar gallons (zie subparagraaf I):



1.6 Contactloos betalen: computerprogramma's

Heb je net zoals ik het tekentalent van een kleuter, beste lezer? Dan kun je grafieken (evenals tabellen, trouwens) ook door een computer laten maken. In de online leeromgeving van *Statistiek in stijl* bespreek ik hoe dat werkt met twee populaire programma's: Excel en SPSS. Aan jou (of je docent) om te beslissen welke je gebruikt.

I Microsoft Office Excel
Welke computer heeft er geen Excel? Dit programma is geknipt voor het maken van overzichtelijke tabellen en grafieken. In de online leeromgeving neem ik een paar van de meest toegankelijke faciliteiten door: *Draaigrafiek en draaitabel*, en *Gegevensanalyse*. Vooral met die eerste moet je gewoon een tijdje fröbelen, dan word je er vast vanzelf handig mee.

II IBM SPSS Statistics
Het programma SPSS *(Statistical Package for the Social Sciences)* is speciaal ontworpen voor statistische analyses. Ermee leren omgaan kost wat meer oefening, maar dan heb je ook wat. In de online leeromgeving maak ik je wegwijs in de *Chart Builder* en in drie nuttige algoritmen: *Frequencies, Descriptives* en *Explore*.

Wil je met mij meedoen? Ga dan naar *statistiekinstijl.noordhoff.nl*, log in bij de online leeromgeving en download de instructies en gegevensbestanden voor hoofdstuk 1.

Samenvatting

Variabelen en meetniveaus

Algemeen	Specifiek	Wat is dat nou weer?
Categorisch	**Nominaal:** *benoeming*	Gelijkwaardige categorieën, bijvoorbeeld geslacht (man/vrouw). Bijzonder geval: *dichotoom* (slechts twee categorieën).
	Ordinaal: *ordening*	Geordende categorieën, bijvoorbeeld opleiding (laag/midden/hoog).
Kwantitatief	**Interval:** *afstand*	Getallen met betekenis, bijvoorbeeld IQ (50-150). Dat komt doordat de afstand tussen opeenvolgende eenheden altijd even groot is (in principe).
	Ratio: *verhouding*	Een intervalvariabele met een absoluut nulpunt, bijvoorbeeld leeftijd. Het onderscheid tussen ratio en interval is niet van cruciaal belang.

Meet-niveau	Geschikte tabellen	Geschikte grafieken	Geschikte centrummaten	Geschikte spreidingsmaten
Nominaal	Frequentietabel	Cirkeldiagram (alleen nominaal) **Staafdiagram** Stapeldiagram	Modus Modus Mediaan	*Geen*
Ordinaal				
Interval	Stamdiagram	**Histogram** Naalddiagram Frequentiepolygoon Boxplot	Modus Mediaan Gemiddelde	Interkwartielafstand (IQR) Standaardafwijking*
Ratio				

* Indien je als centrummaat de mediaan rapporteert, hoort daarbij als spreidingsmaat de interkwartielafstand (let op: de IQR is alleen mogelijk bij kwantitatieve variabelen). Bij het gemiddelde hoort de standaardafwijking.

Centrummaten, spreidingsmaten et cetera

De *i* in sommige van deze formules staat voor de proefpersoon.

Modus **Centrummaat: de score die het vaakst voorkomt.**	
Berekening	Niet van toepassing. Spoor gewoon de meest voorkomende score op.
Gebruik	Kan altijd. Vertelt je iets over de *frequentie* van scores. Bij tweetoppige verdelingen niet één maar twee modi rapporteren.
Mediaan **Centrummaat: de middelste score (50% ligt eronder en 50% erboven).**	
Berekening	Orden de scores van laag naar hoog en vind de middelste met *rang* $= \frac{N+1}{2}$. Krijg je een geheel getal (bijvoorbeeld 6), dan had je een oneven aantal scores. Zoek de score op de plaats van het getal (hier de 6e). Krijg je een kommagetal (zeg 6,5), dan had je een even aantal scores. Neem het gemiddelde van de 6e en de 7e score (in dit geval).
Gebruik	Kan vanaf een ordinale variabele. Vertelt je iets over de *frequentie* en de *ordening* van scores. Ook bruikbaar bij scheve verdelingen en uitschieters (is resistent). Rapporteer bij tweetoppige verdelingen niet één maar twee medianen: één per subgroep.

(Rekenkundig) gemiddelde **Centrummaat: het zwaartepunt van de scores.**	
Berekening	$\bar{X} = \frac{\sum X_i}{N}$ Oftewel: tel alle scores op en deel ze door het totale aantal scores.
Gebruik	Kan vanaf een intervalvariabele. Vertelt je iets over de *frequentie*, de *ordening* en de *waarde* van scores. Wees voorzichtig met gebruik bij scheve verdelingen of uitschieters (is niet resistent). Rapporteer bij tweetoppige verdelingen één gemiddelde per subgroep.
Standaardafwijking/standaarddeviatie **Spreidingsmaat: de gemiddelde afwijking van het gemiddelde.** **Het kwadraat hiervan (s_x^2), berekend in stap 2, noemen we de variantie.**	
Berekening	$s_X = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N - 1}}$ 1) Neem van elke score het verschil met het gemiddelde, kwadrateer deze en tel ze op. 2) Deel de uitkomst door het totale aantal scores minus 1. 3) Trek hieruit de vierkantswortel.
Gebruik	Kan vanaf een intervalvariabele. Wees voorzichtig met gebruik bij scheve verdelingen of uitschieters (is niet resistent). Rapporteer bij tweetoppige verdelingen één standaardafwijking per subgroep.
Kwartielen **Q_1 is het eerste kwartiel (ook wel het 25e percentiel): het scheidt de laagste 25% van de scores van de hoogste 75%.** **Q_2 is het tweede kwartiel (ook wel het 50e percentiel): het scheidt de laagste 50% van de scores van de hoogste 50%.** **Q_3 is het derde kwartiel (ook wel het 75e percentiel): het scheidt de laagste 75% van de scores van de hoogste 25%.**	
Berekening	Orden eerst de scores van laag naar hoog en zoek de 'gewone' mediaan. Dit is in feite Q_2. Zoek daarna de mediaan van de waarden *onder* Q_2: dit is Q_1. De mediaan van de waarden *boven* Q_2 – je raadt het al – is Q_3.
Gebruik	Kan vanaf een ordinale variabele.
Interkwartielafstand (IQR) **Spreidingsmaat: het bereik van de 50% middelste scores.**	
Berekening	$IQR = Q_3 - Q_1$
Gebruik	Kan vanaf een intervalvariabele (!). Ook bruikbaar bij scheve verdelingen en uitschieters (is resistent).
Vijfgetallensamenvatting **Vijf grootheden op een rij die de complete verdeling samenvatten: minimum, Q_1, mediaan, Q_3 en maximum.**	
Berekening	Zoals hierboven (voor de kwartielen). Uitschieters hoeven niet mee te tellen als minimum of maximum (maar mogen dit wel).
Gebruik	Kan vanaf een intervalvariabele.
1,5×IQR-criterium **Criterium om een score als uitschieter te identificeren. (Ook andere criteria zijn mogelijk.)**	
Berekening	Bereken de volgende grenzen: $Q_1 - (1{,}5 \times IQR)$ en $Q_3 + (1{,}5 \times IQR)$ Als een score buiten dit bereik valt, beschouwen we hem als een uitschieter.
Gebruik	Kan vanaf een intervalvariabele.

Lineaire transformaties

In deze formules staat *i* voor de proefpersoon.

Naam	Neem me niet kwalijk?	Berekening
Multipliceren	Alle scores vermenigvuldigen met een bepaald getal.	$X_i' = m \times X_i$ (*m* is een getal, bijvoorbeeld 10)
Centreren	Alle scores zo opschuiven dat het gemiddelde 0 wordt.	$C_i = X_i - \bar{X}$
Standaardiseren	Centreren en vervolgens multiplice-ren: schuif het gemiddelde op 0 en verander daarna de standaard-afwijking in 1.	$z_i = \frac{X_i - \bar{X}}{s_X}$

Hoe nu verder?

Cursus voor cupcakes
Hoofdstuk 2:
Categorische verbanden

1A Opdrachten voor perziken

De meeste mensen kunnen zich tegenwoordig geen leven meer voorstellen zonder smartphone, maar het is verstandig om je bewust te zijn van waar en hoeveel je hem gebruikt. In 2016 verrichtte telecomleverancier Fony een kleinschalig onderzoek naar buitensporig en onveilig smartphonegebruik. 16 proefpersonen vulden een online enquête in (op hun telefoons …) die bestond uit een breed scala aan vragen. De volgende opdrachten presenteren je een selectie van de resultaten. Slaag jij erin alles correct te interpreteren?

Opdracht 1

Als eerste kreeg iedere respondent het verzoek zijn smartphonemodel te bevestigen (het enquêteprogramma detecteerde dit automatisch), zoals (ik heb de modelnamen aangepast om sluikreclame te voorkomen) een yPhoon 6, een Simsong Milky Way S, een RaspBerry Z50 et cetera. Sommige apparaten waren moderner en krachtiger dan andere, dus om ze te vergelijken, classificeerde Fony elke smartphone op een schaal van 1 tot 5 (een zogenaamde Likertschaal):

Likertschaal

- ❑ 1 (heel goedkoop en simpel)
- ❑ 2 (tamelijk goedkoop en simpel)
- ❑ 3 (middensegment)
- ❑ 4 (tamelijk duur en geavanceerd)
- ❑ 5 (heel duur en geavanceerd)

Het meetniveau van de Likertschaal wordt als *ordinaal* beschouwd. Kijk nu naar drie proefpersonen. Respondent Nout had een telefoon die 1 punt waard was, Liza bezat een telefoon ter waarde van 2 punten, en de telefoon van Yoeri was 3 punten waard. Dit houdt in dat:

a Yoeri's smartphone drie keer zo duur was als die van Nout
b Liza's smartphone evenveel verschilde van die van Nout als die van Yoeri
c Nout qua smartphonemodel meer verschilde van Yoeri dan van Liza
d De bovenstaande alternatieven zijn allemaal juist

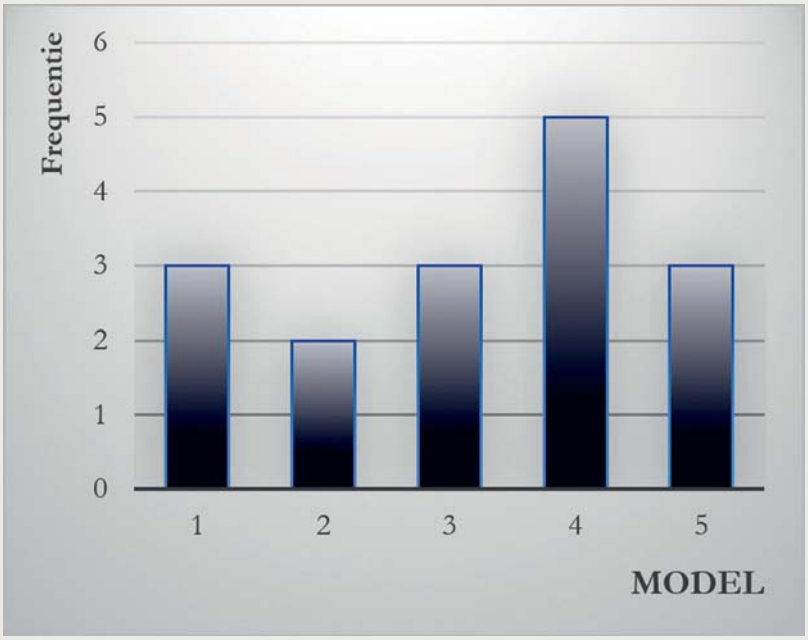
Opdracht 2

Figuur 1.22 geeft de scores weer op de modelschaal van opdracht 1. Zoals eerder aangegeven is $N = 16$.

2.1 Lijkt het je gepast een modus te bepalen? Zo ja, hoe groot is deze?
2.2 Kunnen we een mediaan rapporteren? Zo ja, welke waarde heeft die? Verdient hij de voorkeur boven de modus?
2.3 Zou je een gemiddelde uitrekenen?

2.4 En een standaarddeviatie?
2.5 Een interkwartielafstand misschien?

FIGUUR 1.22 Staafdiagram van de telefoonmodellen

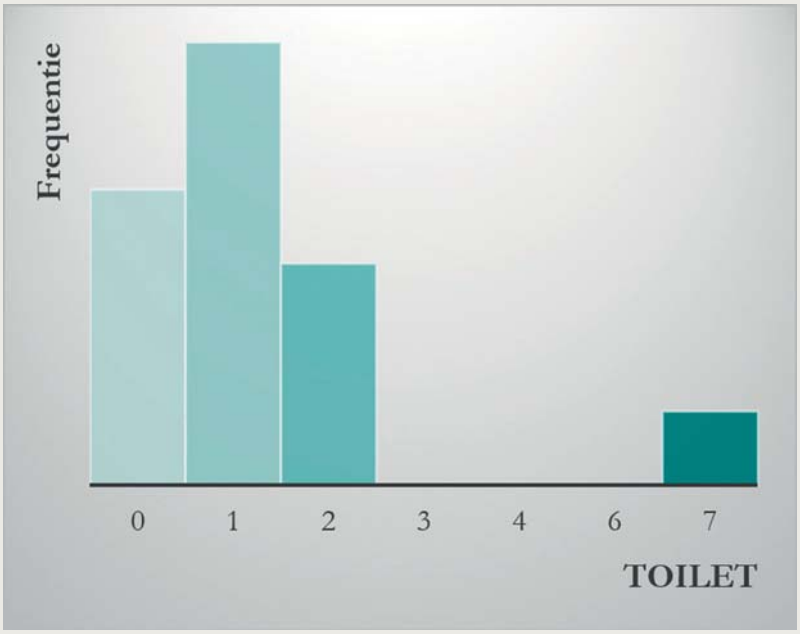


Opdracht 3
De meeste respondenten gebruikten hun smartphone overal, ook op de wc, zelfs terwijl ze hun behoefte deden. De enquête van Fony vroeg hun hoe vaak in hun leven ze per ongeluk hun telefoon in het toilet hadden laten vallen. Het bedrijf had al enige verwachtingen over de uitkomst. Jij ook?

3.1 Verwacht je dat de verdeling van de valpartijen scheef naar links zal zijn, symmetrisch of scheef naar rechts? Probeer je het histogram voor te stellen en interpreteer het.
3.2 Verwacht je een eentoppige verdeling of denk je dat ze uit meerdere subgroepen bestaat?
3.3 Wat denk je dat we het best kunnen doen als we een onderzoek opstarten: wachten tot de statistiek ons de verdeling vertelt, of hierover van tevoren theorieën opstellen?

Opdracht 4

FIGUUR 1.23 Histogram van het aantal toiletincidenten



Figuur 1.23 toont de verdeling van de variabele uit de vorige opdracht, het aantal keren dat iemands telefoon in het toilet belandde. Het bleek dat de gemiddelde smartphone 1,36 keer een duik had genomen ($\overline{X}$ = 1,36). Volgens de meeste respondenten gebeurde het ongelukje toen ze dronken waren, hetgeen wellicht de hoge scores verklaart. De exacte frequenties staan niet op de y-as. Desondanks maakt deze grafiek duidelijk dat de mediaan waarschijnlijk

a Kleiner dan 1,36 was
b Gelijk aan 1,36 was
c Groter dan 1,36 was

Opdracht 5

25% van de respondenten had zijn of haar telefoon 0 keer in het toilet laten vallen. 25% had hem juist 2 keer of vaker laten vallen. Zoals je ziet, was de telefoon van de meest onvoorzichtige persoon maar liefst 7 keer uit diens handen geglipt. Was deze persoon een uitschieter volgens het 1,5 × IQR-criterium?

a Ja
b Nee
c We hebben de kwartielen nodig om dit criterium toe te passen en die zijn onvindbaar 😐

Opdracht 6

Velen gebruiken hun telefoons af en toe te voet of op de fiets. Hoe vaak hadden de proefpersonen bijna een ongeluk veroorzaakt door zonder te kijken de straat over te steken? Figuur 1.24 en 1.25 laten de verdeling van deze variabele zien. Als je wilt, kun je nagaan dat:

- de modus gelijk was aan 0, de mediaan gelijk aan 1, en het gemiddelde ook gelijk aan 1;
- de interkwartielafstand 1,5 was en de standaardafwijking 1,32.

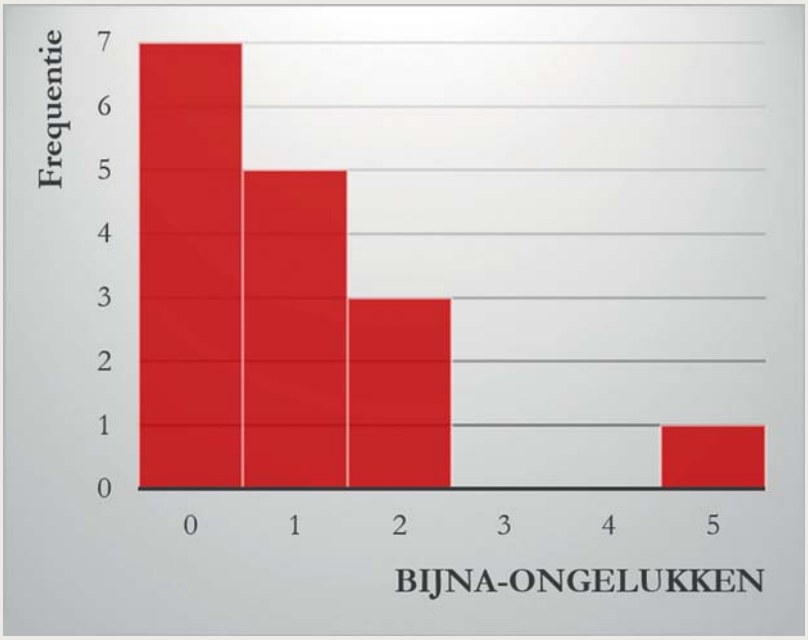
Kort daarop stuurde een zeventiende proefpersoon haar enquête op naar Fony. Het bleek dat zij één keer bijna het loodje had gelegd door onoplettend de straat over te steken (haar score was dus 1).

6.1 Bespreek voor elk van de centrummaten (modus, mediaan en gemiddelde) of ze veranderden door deze extra respondent. Zo ja, wat gebeurde ermee?

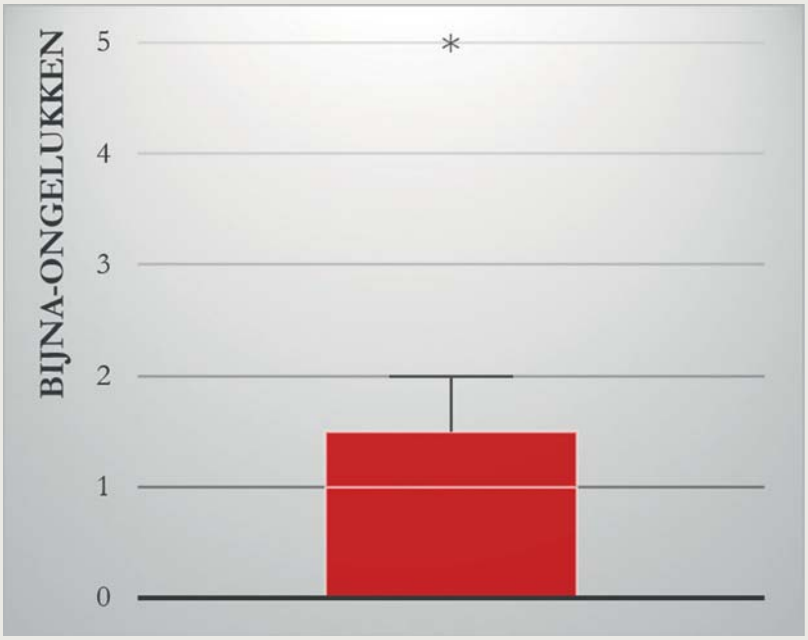
6.2 Lukt dit je ook voor de standaardafwijking?

6.3 Geef voor de verandering van de standaardafwijking een wiskundige verklaring. Kijk naar de formule en praat zowel over de kwadratensom (teller) als over $N - 1$ (noemer).

FIGUUR 1.24 Histogram van het aantal bijna-ongelukken



FIGUUR 1.25 Boxplot van het aantal bijna-ongelukken



Opdracht 7

Fony transformeerde het aantal keren dat de respondenten bijna een verkeersongeluk veroorzaakt hadden naar z-scores. Beschouw twee proefpersonen: Nout had een z-score van –0,76, maar Yoeri had er een van +0,76. Wat kunnen we afleiden uit deze z-scores?

a Yoeri had zich op de weg vaker laten afleiden dan Nout
b Yoeri en Nout zaten allebei dicht bij het gemiddelde
c Yoeri en Nout bevonden zich even ver van het gemiddelde
d De bovenstaande alternatieven zijn allemaal juist

Opdracht 8

In haar onderzoek heeft Fony diverse variabelen in kaart gebracht: het smartphonemodel van de respondent, hoe vaak hij zijn telefoon in het toilet heeft laten vallen, en hoe vaak hij tijdens het appen bijna onder een vrachtwagen is gelopen. Zou je nog meer willen weten over deze variabelen, bijvoorbeeld of er een zeker verband tussen hen bestaat?

1B Opdrachten voor piraten

De afgelopen jaren heeft de bankensector in South Corell zich nog minder populair gemaakt dan ze in de publieke opinie al was. Fraude, nationaliseringen en bestuurders die vertrokken met enorme bonussen; menige burger koesterde wraakgevoelens. Midden in de economische crisis speelde wetenschapper Nguyen op deze emoties in. Hij nodigde een groot aantal willekeurige proefpersonen uit voor een bezoek aan de investeringsbank Manly Brothers, en legde ze een document voor met informatie over de vele schandalen waarin de bank het afgelopen jaar verzeild was geraakt. Na het lezen mocht iedere proefpersoon een kopje koffie gaan halen bij de receptie van Manly Brothers. De receptioniste gaf hem of haar de koffie, bedankte de deelnemer voor zijn deelname en liep vervolgens weg, maar ze liet met opzet een kluisje naast de balie openstaan dat een half miljoen dollar aan bankbiljetten bevatte. Nguyen keek vervolgens hoeveel geld de proefpersoon uit het kluisje stal. Iedere dief werd meteen gearresteerd. Overigens had Nguyen, als voorzitter van de ethische commissie van zijn universiteit, zijn onderzoek van tevoren netjes aan de commissie voorgelegd en het persoonlijk goedgekeurd.
De volgende vragen hebben betrekking op de data die hij heeft verzameld.

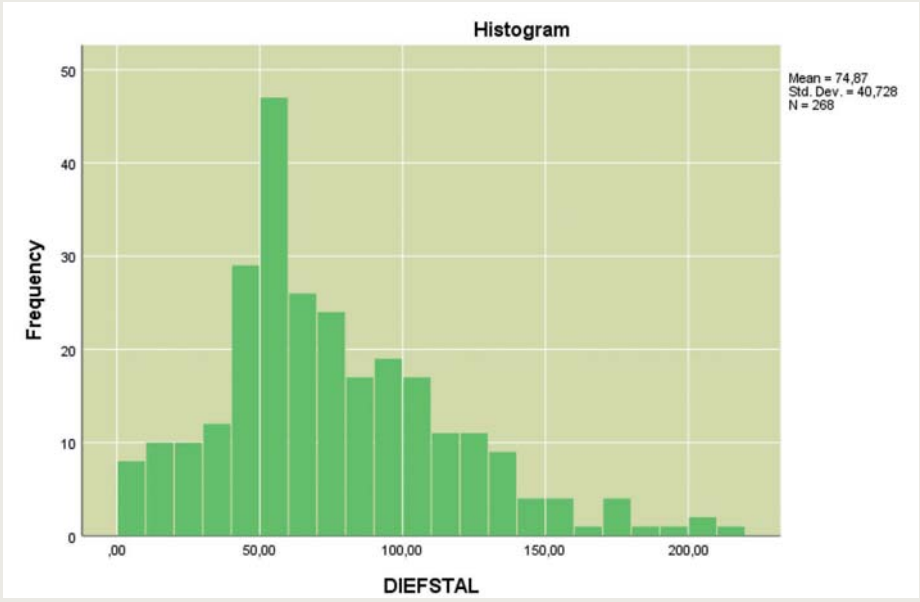
Opdracht 1

Figuur 1.26 laat zien hoe het aantal gestolen dollars was verdeeld onder de proefpersonen, in duizendtallen (het gerapporteerde gemiddelde staat dus voor $74.870).

1.1 Vooraf verwachtten Nguyen en zijn collega Ace dat deze scoreverdeling scheef naar rechts zou zijn. Hun voorspelling klopte, zo blijkt. Wat denk je dat hun argument voor deze scheefheid was?

1.2 Bij het zien van het histogram merkte Ace op dat de verdeling toch minder scheef was dan hij had gehoopt. Wat zal hij daarmee bedoeld hebben?

FIGUUR 1.26 Histogram van de diefstalvariabele



Opdracht 2

2.2 Werp een blik op de grootheden naast het histogram. Jane M. heeft 35 (× 1000) dollar gestolen. Wat is haar z-score op de diefstalvariabele?

2.3 Henry S. heeft 50 (×1000) dollar gestolen. Stel dat we dit bedrag eerst uitdrukken in *euro's*. Vervolgens transformeren we het naar een z-score. Heeft Henry S. dan een hogere of lagere z-score dan Jane M.?

Opdracht 3

Dit is de vijfgetallensamenvatting van de diefstalvariabele, inclusief potentiële uitschieters:

0 49 67 100 215

3.1 Gebruik het 1,5 × IQR-criterium. Bevat de dataset uitschieters?

3.2 Kijk nog eens naar het histogram. Ben je nog steeds van mening dat er uitschieters zijn?

Opdracht 4

Uit de politierapporten bleek dat één deelnemer minderjarig was. Nguyen vond dat zijn onderzoek alleen hoorde te gaan over Corelliaanse burgers van 18 jaar of ouder, en hij besloot deze scholier buiten beschouwing te laten voor zijn wetenschappelijke artikel. Zij had 200 (×1000) dollar gejat. Aanvankelijk was de mediaan gelijk aan 67. Verrassend genoeg bleef deze hetzelfde toen de scholier uit de steekproef werd verwijderd. Leg uit hoe dit kon gebeuren:

a Het meisje was een uitschieter, en de mediaan is resistent tegen uitschieters

b In de complete dataset hadden de middelste twee personen hetzelfde geldbedrag gestolen

c De mediaan was gelijk aan de modus, die niet veranderde

Opdracht 5

Alle proefpersonen die geld uit de kluis hadden gestolen, werden geregistreerd op het politiebureau en gesorteerd op de beginletter van hun

achternaam. De map waarin hun dossier terechtkwam (A tot Z) is een variabele. Onderzoeker Ace merkte op dat het meetniveau van deze variabele nominaal is. Nguyen was het daar niet mee eens en beweerde dat het ordinaal is, aangezien de letters altijd gesorteerd worden in dezelfde volgorde. Hun collega King beweerde zelfs dat het interval is, want A is de 1e letter van het alfabet, B is de 2e, enzovoort, helemaal tot aan de 26e. Wie had de waarheid aan zijn zijde?

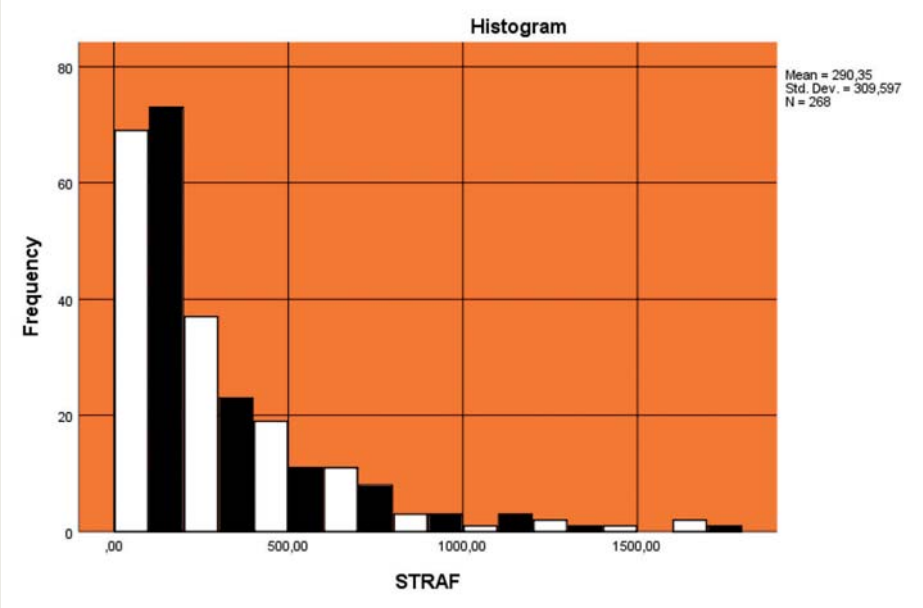
Opdracht 6

De rechter van de lokale rechtbank had er geen zin in 268 aparte zaken te houden, en besloot één algemeen oordeel te vellen over alle deelnemers, afhankelijk van het geldbedrag dat ze hadden gestolen. Hij keek naar dit bedrag (in duizendtallen), deelde het door 5, en kwadrateerde het. De uitkomst was het aantal dagen dat iedere proefpersoon naar de gevangenis werd gestuurd. Dus:

$$STRAF = \left(\frac{DIEFSTAL}{5}\right)^2$$

Het histogram (figuur 1.27) toont je hoe deze nieuwe strafvariabele was verdeeld. Vergelijk het met het eerdere histogram van de diefstalvariabele. Vormde de formule hierboven een *lineaire* transformatie?

FIGUUR 1.27 Histogram van de strafvariabele



a Ja, want de verdeling bleef scheef naar rechts
b Nee, want het gemiddelde en de standaardafwijking veranderden niet in dezelfde mate
c Nee, want de vorm van de verdeling veranderde

De uitwerkingen vind je op de website, statistiekinstijl.noordhoff.nl.