3. Beschrijvende statistiek

Beschrijvende statistiek dient om een overzicht te creëren van de data en deze samen te vatten. Er moet onderscheid gemaakt worden tussen kwantitatieve en categorische data. Bij deze typen data kunnen niet altijd dezelfde beschrijvende statistieken gebruikt worden.

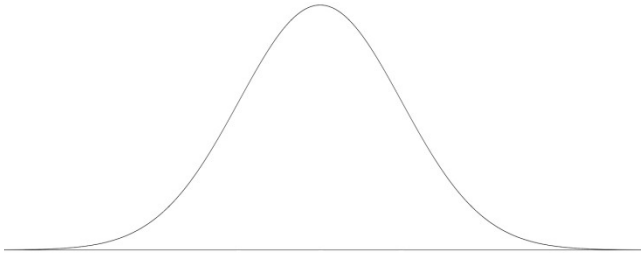
Bij categorische data is het voor het overzicht het makkelijkst als de categorieën in een lijst staan met daarbij de frequenties achter iedere categorie. Om de categorieën beter met elkaar te kunnen vergelijken worden vaak ook de relatieve frequenties weergegeven. De relatieve frequentie van een categorie is het percentage van de totale observaties die in die categorie vallen: het totale aantal observaties in die categorie, gedeeld door het totale aantal observaties * 100. Er kunnen ook proporties gebruikt worden. Dat gaat op dezelfde manier, maar dan vermenigvuldig je niet met 100. De som van alle proporties moet uiteindelijk 1.00 zijn, en de som van alle percentages moet 100 zijn.

Voorbeeld (relatieve) frequentietabel:

	Frequentie	Proportie	Percentage
Man	150	0.43	43%
Vrouw	200	0.57	57%
Totaal	350 (=n)	1.00	100%

Naast tabellen wordt ook vaak gebruik gemaakt van meer visuele weergaven, zoals staafdiagrammen, taartdiagrammen en histogrammen. Wanneer deze worden toegepast op een populatie spreekt men van een populatiedistributie; wanneer ze worden toegepast op een steekproef is dat een steekproefdistributie.

De normale distributie



Een normale distributie heeft een veronderstelde bel-vorm (zie links). Deze is symmetrisch. De twee uitersten worden staarten genoemd (*tails*). Wanneer de ene staart langer is dan de andere, en de verdeling dus niet symmetrisch, is de verdeling linksscheef of rechtsscheef (*skewed*).

Centrummaten

Centrummaten geven een idee over waar het midden van de data ligt. De meest bekende is het gemiddelde: de som van de observaties gedeeld door de totale hoeveelheid observaties. Bijvoorbeeld: een variabele (*y*) heeft de waarden 34 (*y1*), 55 (*y2*) en 64 (*y3*). Het gemiddelde ($\bar{y}$) is (34 + 55 + 64)/3 = 51. De berekening van het gemiddelde ziet er in een formule als volgt uit: $\bar{y} = \frac{\sum y_i}{n}$ (waarbij i = 1 tot n).

Het gemiddelde kan alleen gebruikt worden bij kwantitatieve data en is zeer gevoelig voor uitschieters (*outliers*). Wanneer meerdere gemiddeldes worden berekend, kan je die noteren als $\bar{y}$ 1 en $\bar{y}$2.

Een tweede centrummaat is de mediaan. De mediaan is de middelste observatie. Bijvoorbeeld: een variabele heeft de waarden 1, 3, 8 en 10. De mediaan is dan (3 + 8)/2 = 5,5. Echter, wanneer de variabele de waarden 1, 3, 5, 8 en 10 heeft, dan is de mediaan 5.

Behalve kwantitatieve data is de mediaan ook geschikt voor categorische data, zij het dat deze van ordinaal meetniveau moet zijn. Bij volledige symmetrische data zouden de mediaan en het gemiddelde hetzelfde moeten zijn. Bij een scheve verdeling ligt het gemiddelde, ten opzichte van de mediaan, dichter bij de staart. Tot slot is de mediaan niet gevoelig voor uitschieters. Dit is zowel

iets positiefs als iets negatiefs. Het is positief, want als er één uitschieter in de data zit, de mediaan geen vertekend beeld geeft. Maar het is ook negatief, want variabelen kunnen er van elkaar variëren met een enorme spreiding, terwijl de mediaan dan soms dezelfde middenwaarde aangeeft.

Een derde maat is de modus: de waarde die het vaakst voorkomt. Deze is het nuttigst bij discrete variabelen en dus categorische data, maar kan in principe voor alle typen gebruikt worden.

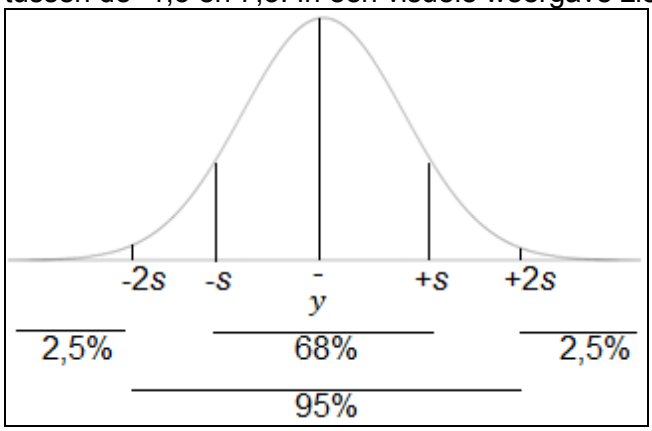
Variabiliteit
Naast het gebruik van centrummaten is het goed om ook de spreiding van de data te beschrijven. Je beschrijft dan de variabiliteit van de waardes van een variabele uit de data, bijvoorbeeld de spreiding van het inkomen van de respondenten.

Ten eerste kan hierbij het bereik (*range*) worden vermeld: het verschil tussen de laagste en de hoogste observatie. Bijvoorbeeld: de waarden 4, 10, 16 en 20. Het bereik is 20 – 4 = 16.

Ten tweede gebruikt men de standaarddeviatie (*s*): het verschil tussen een gemeten waarde (*yi*) en het gemiddelde ($\bar{y}$). Elke observatie heeft zijn eigen standaarddeviatie. Deze kan zowel positief als negatief zijn. Hij is positief wanneer de observatie een hogere waarde heeft dan het gemiddelde, en negatief wanneer deze een lagere waarde heeft dan het gemiddelde. Behalve dat je dit voor iedere observatie apart kan doen, kan je ook de standaarddeviatie van een variabele berekenen, door de som te nemen van alle losse standaarddeviaties. Hierbij hoort de volgende formule: $s = \sqrt{\frac{\Sigma(yi-\bar{y})^2}{n-1}}$. Het bovenste gedeelte $\Sigma(yi-\bar{y})^2$ wordt som van de kwadraten (*sum of squares*) genoemd. Dit gedeelte is belangrijk: het kwadrateert de afzonderlijke standaarddeviaties van de observaties, zodat de som ervan altijd positief is.

Hoe groter *s*, hoe groter de spreiding van waardes van de variabele. En: als *s* = 0, dan is er dus helemaal geen variabiliteit in de data.

Interpretatie van s
Er zijn drie vuistregels met betrekking tot de interpretatie van *s*. Ten eerste ligt 68% van de data tussen $\bar{y}$ – *s* en $\bar{y}$ + *s*. Ten tweede ligt 95% tussen de $\bar{y}$ – 2*s* en $\bar{y}$ + 2*s*. Ten derde vallen vrijwel alle observaties tussen $\bar{y}$ – 3*s* en $\bar{y}$ + 3*s*. In een voorbeeld: stel $\bar{y}$ = 3 en *s* = 1,5. Dan valt 68% tussen 1,5 en 4,5. Dan valt 95% van de observaties tussen 0 en 6. En vrijwel alle observaties liggen tussen de -1,5 en 7,5. In een visuele weergave ziet dit er als volgt uit:

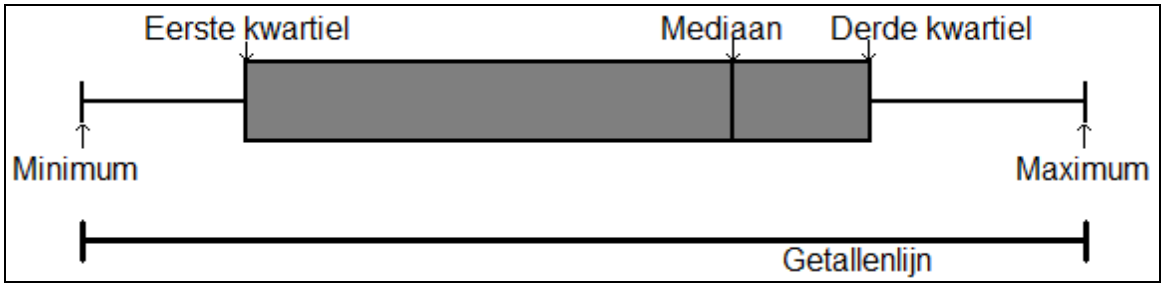


Notatie
Het is belangrijk om bij het noteren van deze statistieken te kijken of het om de steekproef of de populatie gaat. Bij de steekproef is $\bar{y}$ het symbool voor het gemiddelde en *s* het symbool voor de

standaarddeviatie. Bij de populatie is μ het symbool voor het gemiddelde en σ het symbool voor de standaarddeviatie.

Kwartielen
Een distributie kan ook worden omschreven aan de hand van kwartielen. Het “x”-ste kwartiel is het punt waar “x”% van de observaties onder of op valt, en (100 – “x”)% valt erboven. Men onderscheidt drie kwartielen die samen de data in vieren delen. Het eerste kwartiel is “x” = 25. Het tweede kwartiel is “x” = 50, ook wel de mediaan. Het derde kwartiel is “x” = 75. Het verschil tussen het eerste en derde kwartiel wordt interkwartielafstand genoemd. Er kan ook wel gesproken worden van percentielen in plaats van kwartielen.

Boxplots
Al deze beschrijvende gegevens (de mediaan, de kwartielen, minimum en maximum) kunnen grafisch worden weergegeven in een boxplot (zie plaatje). Hiermee wordt een duidelijk overzicht gegeven van de spreiding van de data.



Wanneer er in de data een uitschieter voorkomt wordt deze vaak niet als minimum of maximum aangehouden, maar wordt die met een stip aangegeven buiten de boxplot om. Een observatie wordt als uitschieter beschouwd wanneer deze zich meer dan 1,5 interkwartielafstand van het eerste of derde kwartiel bevind.