Hoofdstuk 1: Distributies

Inleiding
Statistiek is de wetenschap van kennis opdoen op basis van data. Dit hoofdstuk gaat over de verschillende soorten data die we kunnen verzamelen en hoe datasets georganiseerd zijn. Ook wordt aandacht besteed aan het verwerken van data door naar grafieken te kijken. Daarna kijken we naar het proces van leren van data door numerieke samenvattingen te berekenen. Tot slot maken we de overstap van datasamenvattingen naar statistische modellen. Hier wordt de *normaalverdeling* geïntroduceerd. Deze verdelingen spelen een cruciale rol in de methoden die men gebruikt voor het trekken van conclusies uit verschillende datasets.

1.1 Data
Data bestaat uit numerieke waarden. Statistische analyse start met een dataset.
Een dataset wordt geconstrueerd door te bepalen welke *cases* (of *units*) we willen bestuderen. Voor elke case verzamelen we informatie over eigenschappen die *variabelen* genoemd worden.

- *Cases* zijn de objecten die beschreven worden door een dataset. Dit kunnen klanten, bedrijven, proefpersonen of andere objecten zijn.
- Een *label* is een speciale variabele die gebruikt wordt in sommige datasets om verschillende cases van elkaar te onderscheiden.
- Een *variabele* is een eigenschap van een case.
- Verschillende cases kunnen verschillende *waarden* hebben op de variabelen.
- Een *categorische* variabele plaatst een individu in één of van de twee of meer groepen of categorieën. Een voorbeeld is sekse.
- Een *kwantitatieve* variabele heeft numerieke waarden waarmee gerekend kan worden. Een voorbeeld is lengte: iemand van twee meter is twee keer zo lang als iemand van één meter.
- Een *distributie* van een variabele vertelt ons welke waarden van een variabele bij individuen voorkomen en hoe vaak deze waarden voorkomen.

We gebruiken de term *units of measurement* (meeteenheden) om te verwijzen naar de manier waarop een variabele gemeten wordt. Tijd wordt bijvoorbeeld in uren, minuten of seconden gemeten, de lengte van een kind in meters of centimeters. Deze meeteenheden zijn een belangrijk deel van de beschrijving van een kwantitatieve variabele.

De belangrijkste eigenschappen van een dataset
Bij elke dataset hoort bepaalde achtergrondinformatie die helpt bij het interpreteren van de data. Denk hierbij aan de volgende punten:

1. *Wie*? Welke cases beschrijven de data? *Hoe veel* van deze cases bevat de dataset?
2. *Wat*? Hoe veel *variabelen* bevat de data? Wat zijn de *precieze definities* van die variabelen? Wat zijn de meeteenheden voor elke kwantitatieve variabele?
3. *Waarom*? *Welk doel* hebben de data? Hopen we een specifieke vraag te kunnen beantwoorden? Willen we conclusies trekken over cases waarover we geen data hebben? Zijn de gebruikte variabelen geschikt voor het beoogde doel?

Voor het verwerken van de data is het goed om een *spreadsheet* te gebruiken. Dit kan bijvoorbeeld in Excel (zie figuur 1.2 in het boek). Het is belangrijk om bij de variabele-namen spaties te vermijden, omdat deze in sommige statistische software niet toegestaan zijn. In plaats van een spatie kan een underscore (_) gebruikt worden.

Wanneer we een variabele geschikt willen maken om mee te rekenen, kunnen we de variabele *transformeren*. Zo kunnen de letterbeoordelingen uit het Amerikaanse schoolsysteem omgezet worden in cijfers (A=4, B=3, etc.). Dit kan alleen wanneer het verschil tussen A en B even groot is als bijvoorbeeld het verschil tussen C en D.

Een onderdeel van het goed worden in statistiek is weten welke variabelen belangrijk zijn en hoe deze het beste gemeten kunnen worden. Vaak is voor detais van bepaalde metingen kennis nodig van het specifieke studieveld. Zorg er in ieder geval voor dat elke variabele echt meet wat jij wilt dat hij meet. Een slechte keuze van variabelen kan leiden tot misleidende conclusies.

1.2 Distributies grafisch weergeven

Verkennende data-analyse

Verkennende data-analyse (exploratory data analysis) houdt in dat de belangrijkste kenmerken van een dataset worden beschreven. De volgende twee strategieën kunnen in dit verband gebruikt worden:

- Onderzoek elke variabele eerst afzonderlijk. Pas daarna dient gekeken te worden naar de relatie tussen de variabelen.
- Geef grafisch de waarden van variabelen weer. Daarna kunnen numerieke samenvattingen gemaakt worden van deze waarden.

De waarden van een categorische variabele zijn labels voor de categorieën, zoals ‘vrouw’ en ‘man’. De *distributie* van een categorische variabele laat zien hoeveel van de onderzochte mensen een bepaalde waarde heeft gescoord *(count).* Dit kan ook door middel van *percentages* vermeld worden.

Diagrammen voor categorische variabelen

Een distributie kan grafisch weergegeven worden door een:

- *Staafdiagram (bar graph):* De hoogtes van de staven zegt iets over hoe vaak bepaalde waarden voorkomen. De frequenties staan op de y-as en de lengtes van de staven dienen daar dan ook mee te corresponderen.
- *Cirkeldiagram (pie chart*): Hiermee kun je bijvoorbeeld meteen zien of er meer mannen dan vrouwen hebben meegedaan aan een onderzoek. Omdat cirkeldiagrammen geen gebruik maken van schalen, worden hoeveelheden door middel van percentages uitgedrukt. Voor cirkeldiagrammen is het nodig dat alle categorieën, waaruit het geheel bestaat, worden toegevoegd.

Staafdiagrammen zijn makkelijker te interpreteren en zijn ook flexibeler dan cirkeldiagrammen. Ze kunnen allebei gebruikt worden wanneer je wilt dat mensen in één oogopslag kunnen zien hoe het zit met frequenties van waarden van een variabele.

Diagrammen voor kwantitatieve variabelen: Stam-en-bladdiagram (stemplot)
Een stam-en-bladdiagram geeft snel een beeld van de vorm van een distributie, terwijl elke waarde in de oorspronkelijke vorm worden toegevoegd. Zo een diagram is het handigst als er sprake is van niet al te veel observaties (die allemaal groter dan nul zijn). Om een stam-en-bladdiagram te maken, dienen de volgende stappen uitgevoerd te worden:

- Allereerst moet elke waarde opgedeeld worden in een *stam* en een *blad.* De stam is het eerste cijfer en het blad is het laatste cijfer (bij het getal 35 is 3 dus de stam en 5 het blad). Stammen kunnen meerdere cijfers bevatten (bij het getal 135 is 13 de stam), maar een blad bestaat altijd uit maar één cijfer.
- Vervolgens moeten alle stammen onder elkaar genoteerd worden. De kleinste stam moet bovenaan staan. Na dit gedaan te hebben moet een verticale lijn aan de rechterkant van de stammen getrokken worden.
- Tot slot moet het bijbehorende blad in elke rij rechts van de stam genoteerd worden. Er moet met het kleinste blad begonnen worden.

Rug-aan-rugdiagram (back-to-back stemplot)
Een rug-aan-rugdiagram is een variant van de stam-en-bladdiagram. Met zo een diagram kunnen twee gerelateerde distributies vergeleken worden. Zo een diagram maakt gebruik van gemeenschappelijke stammen. Je kunt bijvoorbeeld het gewicht van mannen en vrouwen in een rug-aan-rugdiagram verwerken. De stammen van de gewichten staan dan in het midden en er worden twee lijnen (zowel links als rechts) vanaf de stammen getrokken. Aan de rechterkant kun je dan bijvoorbeeld de bladen van de vrouwen noteren, terwijl je aan de linkerkant de bladen van de mannen opschrijft.

Gevolgen van een grote dataset
Stam-en-bladdiagrammen en rug-aan-rugdiagrammen zijn niet handig wanneer er een grote dataset gebruikt wordt. Het duurt dan erg lang om elke waarde in het diagram te verwerken en dit ziet er bovendien onoverzichtelijk uit. Dit kan echter opgelost worden door het aantal stammen in een diagram te verdubbelen. Dit kan gedaan worden door:

- *Splitting each stem*: Elke stam door twee te delen.
- *Trimming:* Hierbij maak je de cijfers passend wanneer de geobserveerde waarden veel cijfers bevatten. Dit wordt gedaan door de laatste cijfers te verwijderen voordat een stam-en-bladdiagram gemaakt wordt.

Histogrammen
Bij een *histogram* worden de waarden van een variabele opgedeeld in groepen. Daarom worden alleen de frequenties of percentages beschreven die bij de groepen horen. Je mag zelf weten hoeveel groepen je maakt, maar de groepen moeten wel van gelijke grootte zijn. Wel is het belangrijk om te weten dat de manier waarop een histogram eruit ziet kan veranderen wanneer de klassen veranderd worden. Het duurt (in vergelijking tot stam-en-bladdiagrammen) langer om histogrammen handmatig te maken. Ook komen de oorspronkelijke datawaarden niet letterlijk voor in een histogram. Dit is juist wel het geval bij stam-en-bladdiagrammen. Om een histogram te maken moeten drie stappen uitgevoerd worden:

1. Het maken van groepen. Bij een dataset met de IQ-meting van vijftig mensen kun je bijvoorbeeld intervallen maken van 75≤ IQ <85, 85 ≤ IQ <85 enz.
2. Deel de gevonden waarden in per groep. Vervolgens moet geteld worden hoe vaak waarden in een bepaalde groep vallen *(frequenties*). Een tabel met de frequenties die samengaan met elke groep wordt een *frequentietabel* genoemd.
3. Teken tot slot de histogram. Op de horizontale as (X-as) moeten in ons geval de IQ-scores staan, terwijl op de Y-as de frequenties staan. Elke staaf staat voor een groep. Er is geen ruimte tussen de staven, behalve als niemand binnen een bepaalde groep gescoord heeft. Dat is bijvoorbeeld het geval als niemand een IQ-score heeft tussen de 75 en 84.

Verschillen tussen histogrammen en staafdiagrammen
Histogrammen en staafdiagrammen lijken op elkaar, maar zijn niet hetzelfde. Bij een staafdiagram staan de staven niet precies tegen elkaar aan, terwijl dit wel het geval is bij een histogram. Bij een histogram gaat het om de tellingen of percentages van verschillende waarden van een variabele. Een staafdiagram vergelijkt de grootte van verschillende items. De horizontale as van een staafdiagram hoeft geen meetschaal te hebben, maar kan bestaan uit labels. Als men wil weten hoeveel studenten er biologie, psychologie of geneeskunde studeren, dan zijn dit categorische variabelen die je op de X-as kunt zetten. In dit geval dient een staafdiagram gemaakt te worden. Als het gaat om een numerieke variabele (bijvoorbeeld IQ, lengte of gewicht), dan dient een histogram gebruikt te worden.

- Staafdiagrammen zijn dus handig voor *categorische* variabelen, terwijl histogrammen van belang zijn voor *kwantitatieve* variabelen.

Distributies bekijken
Nadat een dataset in een grafiek of diagram verwerkt is, moeten de belangrijkste kenmerken van de distributie onderzocht worden. Het is in dit verband van belang om te kijken naar de volgende zaken:

- Bekijk het *algemene patroon (overall pattern)* en let goed op opvallende afwijkingen van het algemene patroon *(deviations).*
- Ook moet gekeken worden naar de *vorm (shape),* het *midden (center)* en de *spreiding (spread)* binnen een dataset. Het midden van een distributie is de waarde waardoor de helft van de observaties kleiner is dan die waarde en de andere helft groter is dan die waarde. De spreiding van een distributie kan beschreven worden door naar de *kleinste* en *grootste* waarden te kijken. Bij het bekijken van de vorm is het belangrijk of er meerdere pieken in de distributie zijn. Als er sprake is van maar één *piek (mode),* dan noemen we de distributie *unimodaal.* Ook moet bekeken worden of de distributie symmetrisch is of dat er een afwijking naar links of rechts is. Een distributie is *symmetrisch* wanneer de waarden die kleiner en groter zijn dan het middenpunt met elkaar gespiegeld kunnen worden. Als er een *afwijking naar rechts* is *(skewed to the right)*, dan is de rechterstaart (die bestaat uit grotere waarden) veel langer dan de linkerstaart (die uit kleine waarden bestaat). Lengte en IQ zijn variabelen die vaak een (ongeveer) symmetrische distributie hebben. Er zijn maar weinig mensen die extreem klein of extreem groot zijn en het gros van de mensen scoort gemiddeld. Hetzelfde geldt voor IQ-scores. Huizenprijzen hebben een distributie met een afwijking naar rechts. Veel huizen zijn ongeveer even duur, terwijl er een aantal zeer dure villa's bestaat.
- Een belangrijke afwijkende score is een *uitbijter (outlier).* Dit is een individuele score die duidelijk buiten het algemene patroon valt.

Uitbijters
Het vaststellen van uitbijters gaat niet volgens specifieke regels. Het gaat er juist om dat je zelf een mening vormt over welke scores als afwijkend bestempeld moeten worden. Zoek in ieder geval altijd naar waarden die duidelijk anders zijn dan de meeste waarden; het hoeft dus niet alleen te gaan om *extreme* observaties binnen een distributie. Daarnaast is het belangrijk om uitbijters te proberen te verklaren. Een uitbijter kan bijvoorbeeld het gevolg zijn van ongewone omstandigheden.

Tijdplots (time plots)
Wanneer data door de tijd heen verzameld wordt, is het een goed idee om de observaties grafisch op volgorde te verwerken. Het gebruik van histogrammen en stam-en-bladdiagrammen kunnen in dit verband misleidend zijn, omdat er sprake kan zijn van systematische veranderingen door de tijd heen.

- Een *tijdplot (time plot)* van een variabele geeft een grafische weergave van elke observatie in relatie tot het moment waarop deze gemeten variabele is. Tijd moet altijd op de horizontale lijn gezet worden, terwijl de gemeten variabele op de Y-as moet staan. Het verbinden van datapunten (door middel van lijnen) laat zien of er veranderingen door de tijd heen plaatsgevonden hebben. Ook kunnen op deze manier trends ontdekt worden.
- Veel datasets zijn *tijdseries (time series).* Dit zijn metingen van een variabele die op verschillende momenten zijn gedaan. Denk in dit verband bijvoorbeeld aan het meten van de landelijke werkloosheid per kwartaal.
- Een *trend* in een tijdserie is een aanhoudende stijging of daling op lange termijn. Een patroon dat zich in een tijdserie steeds op specifieke momenten herhaalt, wordt *seizoensgerelateerde variatie (seasonal variation)* genoemd. In dat geval wordt *seizoensgerelateerde aanpassing (seasonal adjustment)* uitgevoerd, zodat onderzoeksresultaten geen misleidend effect hebben. Dat het werkloosheidspercentage in december en januari is toegenomen, zegt niet per se dat meer mensen werkloos zijn geworden. Werkloosheidscijfers stijgen namelijk altijd in deze periode, omdat tijdelijke werkkrachten bijvoorbeeld vaak aan het eind van het jaar stoppen met werken. Rekening houden met zo een verschijnsel is een vorm van seizoensgerelateerde aanpassing.

1.3 Distributies met getallen beschrijven

Het gemiddelde *(the mean)*
Een numerieke beschrijving van een distributie begint met een meting van het middenpunt. De meest bekende metingen van het middenpunt zijn het *gemiddelde* en de *mediaan.* Het gemiddelde gaat ook echt om het vinden van de gemiddelde waarde, terwijl de mediaan gaat over het vinden van de middelste waarde.

Om het gemiddelde (mean $\bar{x}$) te vinden moeten alle scores opgeteld worden en worden gedeeld door het aantal scores. Als n aantal mensen de scores x_1, x_2, x_3, x_n hebben, dan is hun gemiddelde:

- $\bar{x} = x_1 + x_2 + x_3 + \ldots + x_n / n$.
- Een andere notatie is: $\bar{x} = 1/n \Sigma x_i$. In deze formule staat Σ als Griekse letter voor ‘alles bij elkaar optellen’.

Het nadeel van het gemiddelde is dat deze maat erg gevoelig is voor de invloed van een aantal extreme observaties. Deze extreme scores kunnen uitbijters zijn, maar dat hoeft niet. Omdat het gemiddelde wordt beïnvloed door extreme scores, zeggen we dat het gemiddelde geen *robuuste maat (resistant measure)* van het middenpunt is. Dat het gemiddelde geen robuuste maat is, blijkt ook uit het feit dat je alleen al één score uit de distributie kunt aanpassen om het gemiddelde te laten veranderen.

De mediaan

De *mediaan M* is het letterlijke middenpunt van en distributie. De helft van de observaties valt onder de mediaan, terwijl de andere helft zich boven de mediaan bevindt. De mediaan van een distributie kan als volgt gevonden worden:

1. Zet alle scores eerst op volgorde (van klein naar groot).
2. Als het aantal observaties oneven is, dan is de mediaan precies het middelste getal. Als er bijvoorbeeld vijf getallen zijn, dan is de mediaan het derde getal. De plaats van de mediaan kan in dit geval als volgt gevonden worden: (n+1)/2. In ons voorbeeld is dat dus: (5+1)/2=3. Deze formule zegt dus niet wat de mediaan is, maar waar de mediaan zich in de reeks getallen bevindt.
3. Als het aantal observaties even is, dan is de mediaan M het gemiddelde van de twee middelste observaties in de distributie. De plaats van de mediaan wordt op dezelfde manier gevonden: M= (n+1)/2.

Het gemiddelde versus de mediaan

Als een distributie helemaal symmetrisch is, dan zijn de mediaan en het gemiddelde hetzelfde. In een distributie die afwijkt naar links of rechts, bevindt het gemiddelde zich meer in de staart dan de mediaan. Dit omdat het gemiddelde veel meer door extreme scores wordt beïnvloed. De staart van een distributie bestaat uit extreme scores.

Spreiding (variabiliteit)

De meest simpele numerieke beschrijving van een distributie moet bestaan uit een maat voor het middenpunt (zoals het gemiddelde en de mediaan), maar ook uit een meting van de spreiding binnen een distributie. We kunnen de spreiding van een distributie beschrijven door verschillende percentielen uit te rekenen. De mediaan deelt de distributie precies in tweeën en daarom zeggen we ook wel dat de mediaan het vijftigste percentiel is. Er is echter nog een *kwartiel* in de bovenste helft van de data. Er is ook een lager kwartiel in de lagere helft van de data. De kwartielen zorgen ervoor dat de data in vieren gedeeld kan worden; elk kwartiel gaat over een kwart van de data. Kwartielen kunnen als volgt berekend worden:

- Eerst moeten alle scores van klein naar groot op volgorde gezet worden. Daarna moet de mediaan voor de hele set berekend worden.
- Het *eerste kwartiel (Q_1)* is de mediaan van de kwart laagste scores van een distributie.
- Het *derde kwartiel (Q_3)* is de mediaan van de kwart hoogste scores een distributie.

Het *p^{ste} percentiel* van een distributie is de waarde waaraan p procent van de scores gelijk is of waar p procent van de scores onder liggen.

Vijf waarden

Om een beschrijving van het middenpunt en de spreiding van een distributie te maken, is het handig om (1) de laagste score, (2) Q_1 , (3) M (de mediaan), (4) Q_3 en (5) de hoogste score te berekenen. Deze waarden worden bij elkaar ook wel de *vijf-getallen-samenvatting* genoemd.

Deze vijf waarden zijn zichtbaar in een boxplot.

- De buitenste twee randen van het doosje (box) in een boxplot staan voor Q_1 en voor Q_3.
- De mediaan wordt weergegeven door de lijn midden in het doosje.
- Twee lijnen (naar boven en naar beneden toe) vanaf het doosje laten zien wat de hoogste waarde is en wat de laagste waarde is.

Interkwartielafstand (interquartile range → IQR)
De bekijken van de grootste en de kleinste waarde zegt in principe weinig over de spreiding binnen de data. De afstand tussen de eerste en het derde kwartiel is een meer robuuste maat voor spreiding. Deze afstand wordt de *interkwartielafstand* genoemd en wordt als volgt berekend:

- IQR: $Q_3 - Q_1$.
- De IQR wordt vaak gebruikt als vuistregel om uitbijters vast te stellen. Vaak wordt een score een uitbijter genoemd als deze 1.5xIQR boven het derde kwartiel of 1.5xIQR onder het eerste kwartiel valt.

Afwijkende distributies
Kwartielen en de IQR worden niet beïnvloed door veranderingen in de staart van een distributie. Ze zijn dus behoorlijk robuust. Wel moet gezegd worden dat geen enkele numerieke waarde van spreiding (zoals de IQR) erg handig is om de spreiding van distributies met een afwijking (naar links of rechts) te beschrijven. De twee kanten van een afwijkende distributie hebben namelijk verschillende spreidingen en dus kan één spreidingswaarde niet toereikend zijn. Een afwijking naar links of rechts kan opgemerkt worden door te bekijken hoe ver het eerste kwartiel en de laagste score afliggen van de mediaan (linker staart) en door te kijken naar hoe ver het derde kwartiel van de hoogste score ligt (rechter staart).

Variantie en standaarddeviatie
Veel vaker dan de *vijf-getallen-samenvatting* wordt de standaarddeviatie (samen met een maat voor het middelpunt) gebruikt om een beeld van een distributie te krijgen. De *standaarddeviatie* meet de spreiding door te kijken naar hoe ver observaties van het gemiddelde af liggen.

- De *variantie (s^2)* van een dataset is het gemiddelde van de gekwadrateerde standaarddeviaties. In formulevorm is dit: $s^2=(x_1-\bar{x})^2+(x_2-\bar{x})^2+ \ldots+(x_n-\bar{x})^2/ n-1$. Een andere juiste formule is: $s^2=1/n-1\ \Sigma(x_i-\bar{x})^2$. In dit verband staat n-1 voor het aantal *vrijheidsgraden (degrees of freedom).*
- Om de *standaarddeviatie (s)* te vinden moet de wortel uit de variantie getrokken worden. Het vinden van de standaarddeviatie is vooral handig als er sprake is van normaalverdelingen. Deze distributies worden in de volgende paragraaf besproken. De standaarddeviatie wordt geprefereerd boven de variantie. Dit omdat het trekken van de wortel uit de variantie ervoor zorgt dat spreiding wordt gemeten volgens de oorspronkelijke schaal van de variabele.

De afwijkingen van het gemiddelde (x_i-$\bar{x}$) laten zien in welke mate scores van het gemiddelde verschillen. Sommige van deze afwijkingen zullen positief zijn, terwijl andere afwijkingen negatief zullen zijn. De som van afwijkingen van de scores zal daarom altijd nul zijn. Om deze reden worden de afwijkingen van het gemiddelde gekwadrateerd; zo komt de berekening namelijk niet uit op nul. De variantie en de standaarddeviatie zullen groot zijn als scores erg verspreid liggen vanaf het gemiddelde. De variantie en de standaarddeviatie zullen klein zijn wanneer de scores dichtbij het gemiddelde liggen.

De standaarddeviatie

- Standaarddeviatie *s* meet de spreiding vanaf het gemiddelde en moet alleen gebruikt worden wanneer het gemiddelde (en dus niet de mediaan) als maat voor middenpunt wordt gekozen.
- De standaarddeviatie is nul wanneer er geen spreiding in een distributie aanwezig is. Dit gebeurt alleen als alle waarden hetzelfde zijn. Als dit niet zo is, dan die standaarddeviatie groter dan nul. Hoe meer spreiding er is, hoe groter s wordt.
- De standaarddeviatie s is, zoals het gemiddelde, niet robuust. De aanwezigheid van een paar uitbijters kan s meteen erg groot maken. De standaarddeviatie is in vergelijking met het gemiddelde zelfs gevoeliger voor extreme scores.
- Distributies met een sterke afwijking (naar links of rechts) hebben grote standaarddeviaties. In dit geval is het niet erg handig om de standaarddeviatie uit te rekenen. De vijf-getallen-samenvatting is vaak handiger dan het gemiddelde en de standaarddeviatie wanneer een afwijkende distributie beschreven moet worden of wanneer een distributie extreme uitbijters heeft. Het gebruik van het gemiddelde en de standaarddeviatie is juist handiger wanneer er weinig uitbijters aanwezig zijn en als de distributie symmetrisch is.

Het veranderen van meeteenheden

Dezelfde variabele kan vaak gemeten worden door middel van verschillende meeteenheden. Temperatuur kan bijvoorbeeld zowel in Fahrenheit als in Celsius gemeten worden. Gelukkig is het gemakkelijk om meeteenheden om te zetten. Dit omdat een verandering in meeteenheid een *lineaire transformatie* van de metingen is. Zo een transformatie verandert de vorm van een distributie niet. Als temperatuurmetingen in Fahrenheit zorgen voor een distributie met aan afwijking naar rechts, dan zal dat zo blijven als de waarden omgezet zijn naar Celsius. De spreiding en het middenpunt zullen echter wel veranderen na zo een verandering. Een lineaire transformatie verandert de oorspronkelijke variabele x in een nieuwe variabele (x_{new}) op basis van de volgende formule:

- $x_{new} = a+bx$. Het toevoegen van de constante a verandert alle waarden van x in dezelfde mate. Zo een aanpassing verandert het nulpunt van een variabele. Vermenigvuldigen met de positieve constante b verandert de grootte van de meeteenheid.
- Om het effect van lineaire transformatie op spreidingsmaten en op maten van het middenpunt te bekijken, is het van belang om elke observatie met het positieve getal b te vermenigvuldigen. Dit zorgt ervoor dat de mediaan, het gemiddelde, de standaarddeviatie en de IQR vermenigvuldigd worden met b.
- Het toevoegen van hetzelfde getal a (of dit getal nou positief of negatief is) aan elke observatie, voegt a toe aan het gemiddelde, de mediaan, de kwartielen en de percentielen. Spreidingsmaten worden echter niet beïnvloed.

1.4 Normaalverdelingen

Dichtheidscurves

Het handmatig maken van histogrammen is onhandig. Tegenwoordig gebruiken wetenschappers dan ook vaak computerprogramma's om histogrammen te maken. Het voordeel van computerprogramma's is dat zij ook een passende curve kunnen maken op basis van de histogram. Dit worden *dichtheidscurves (density curves)* genoemd. Door zo een curve 'vloeit' de histogram als het ware. Gebieden onder de curve staan voor proporties van scores.

- Een dichtheidscurve wordt altijd boven de horizontale as gemaakt.
- Het totale gebied binnen de curve staat gelijk aan 1.
- Een dichtheidscurve beschrijft het algemene patroon van een distributie. Dichtheidscurves kunnen, net zoals distributies, allerlei vormen hebben. Een bijzondere variant is de normaalverdeling, waarbij beide helften van de curve symmetrisch zijn. Uitbijters worden niet beschreven met een dichtheidscurve.

Het meten van het middenpunt en de spreiding bij normaalverdelingen

De *modus* van een distributie beschrijft het piekpunt van de curve. Het gaat dus om de plaats waar de curve het hoogst is. Omdat gebieden onder de curve voor proporties staan, is de mediaan het punt dat precies in het midden ligt.

De *kwartielen* kunnen geschat worden door de curve in ongeveer vier gelijke stukken te verdelen. De IQR is dan de afstand tussen het eerste en het derde kwartiel. Er zijn rekenkundige manieren om de gebieden onder een curve te berekenen. Door deze rekenkundige manieren kunnen we de mediaan en de kwartielen precies berekenen.

Het gemiddelde van een dichtheidscurve is het punt waarop de curve zou balanceren als deze van vast materiaal gemaakt zou zijn. Bij een symmetrische curve liggen de *mediaan* en het *gemiddelde* op hetzelfde punt. Bij een afwijkende distributie is dat niet het geval. Bij een curve met een afwijking naar rechts ligt de mediaan iets meer richting de piek van de curve dan het gemiddelde. Het gemiddelde bevindt zich dus meer naar de staart toe. Bij een afwijkende distributie is het lastig om het balanspunt met het blote oog te bepalen. Er zijn rekenkundige manieren om het gemiddelde en de standaarddeviatie van een dichtheidscurve te berekenen. Kortom:

- De *mediaan* van een dichtheidscurve ligt dus op het punt dat het gebied onder de curve in tweeën deelt.
- Het *gemiddelde* van een dichtheidscurve is het balanspunt waarop de curve zou balanceren als deze van vast materiaal gemaakt zou zijn.
- De mediaan en het gemiddelde zijn hetzelfde voor een symmetrische dichtheidscurve. Het gemiddelde van een afwijkende distributie ligt meer in de richting van de lange staart, terwijl de mediaan meer in de richting van de piek ligt.

Normaalverdelingen

Het gemiddelde van een dichtheidscurve geven we aan met de letter μ. De standaarddeviatie wordt genoteerd aan de hand van het symbool σ. Deze waarden worden benaderd met het steekproefgemiddelde ($\bar{x}$) en de standaarddeviatie (s) die bij deze scores hoort. Normaalverdelingen zijn symmetrisch en unimodaal: ze hebben dus maar één piek. Het veranderen van μ (terwijl de standaarddeviatie onveranderd blijft) zorgt ervoor dat de plaats van de curve op de horizontale as opschuift, terwijl de spreiding hetzelfde blijft. Een curve met een

grotere standaarddeviatie is breder en lager. De standaarddeviatie σ is de spreidingsmaat die bij een normaalverdeling hoort. Samen met μ bepaalt σ de vorm van een normaalverdeling.

Waarom zijn normaalverdelingen belangrijk in de statistiek?

- Normaalverdelingen zijn goede beschrijvingen van distributies die *bij echte data* horen. Het gaat in dit verband om distributies die bijna normaalverdeeld zijn. Voorbeelden zijn distributies van lengte, gewicht en IQ.
- Normaalverdelingen zijn goede benaderingen van de uitkomsten van kansberekeningen, bijvoorbeeld in het geval van het werpen van een munt.
- Tot slot zijn normaalverdelingen handig, omdat statistische berekeningen (die op basis van normaalverdelingen gemaakt zijn), gebruikt kunnen worden voor andere, bijna symmetrische distributies.

Gemeenschappelijke kenmerken
Er zijn veel soorten normaalverdelingen, maar ze hebben een aantal gemeenschappelijke kenmerken. Hieronder worden de belangrijkste kenmerken uiteengezet.

- Ongeveer *68%* van de scores valt binnen 1 standaarddeviatie (σ) van het gemiddelde (μ).
- Ongeveer *95%* van de scores valt binnen twee standaarddeviaties van het gemiddelde.
- Ongeveer *99.7%* van de scores valt binnen drie standaarddeviaties van het gemiddelde.

De bovenste kenmerken staan samen bekend als de *68-95-99.7 regel*. De normaalverdeling met gemiddelde μ en standaarddeviatie σ wordt genoteerd als N(μ,σ). Bij het onderzoek naar de lengte van Nederlandse vrouwen is het bijvoorbeeld mogelijk dat N(1.70,10) wordt gevonden.

Gestandaardiseerde waarden
Als iemand zestig punten op een test heeft gescoord, weet je niet of dit een hoge of lage score is in vergelijking tot alle andere scores. Het is daarom belangrijk om de waarde te standaardiseren.

- Als x een score is uit een distributie met gemiddelde μ en standaarddeviatie σ, dan is de *gestandaardiseerde waarde* van x: $z = (x-\mu)/\sigma$. Een gestandaardiseerde waarde wordt vaak een *z-score* genoemd.
- De gestandaardiseerde waarden van een distributie hebben samen een gemiddelde van 0 en een standaarddeviatie van 1. De gestandaardiseerde normaalverdeling heeft dus de N(0,1) –distributie.

Cumulatieve proporties
Het op precieze wijze berekenen van de proporties onder de normaalverdeling kan door middel van z-tabellen of software.

- Z-tabellen en software berekenen vaak een *cumulatieve proportie*: dit is de proportie observaties in een distributie die onder een bepaalde waarde ligt of daar precies gelijk aan is.

Wanneer een distributie door middel van een dichtheidscurve wordt beschreven, dan is de cumulatieve proportie het gebied onder de curve dat aan de linkerkant van een bepaalde waarde ligt. Hiermee wordt rekening gehouden worden als je bijvoorbeeld juist alleen de proportie wilt hebben dat zich aan de rechterkant van de waarde bevindt. In dat geval moet je 1-de proportie aan de linkerkant berekenen. De z-tabel kan gebruikt worden om proporties onder de curve te achterhalen. Om dit te doen moeten scores wel eerst gestandaardiseerd worden. Een voorbeeld is dat je wilt weten hoeveel studenten minimaal een score van 820 hadden op een bepaalde test. Het gemiddelde blijkt 1026 te zijn en de standaarddeviatie is 209.

- De bijbehorende z-score is: 820-1026/209=-0.99.
- Vervolgens moet de z-tabel gebruikt worden om te kijken welke proportie bij -0.99 hoort. Dat blijkt 0.1611 te zijn. Het gebied rechts van -0.99 is daarom 1-0.1611=0.8389.
- Als je had willen weten hoeveel studenten maximaal een score van 820 hadden behaald, dan was het antwoord 0.1611 geweest.

Normaal kwantielplot (normal quantile plot)
Stam-en-bladdiagrammen en histogrammen word vaak gebruikt om te kijken of een distributie normaal verdeeld is. De *normaal kwantiel plot* is echter de beste grafische manier om normaliteit te ontdekken. Het is niet praktisch om zo een plot zelf te maken. In de meeste gevallen wordt dan ook software gebruikt. Hieronder wordt een algemeen beeld geschetst van hoe zo een plot handmatig gemaakt kan worden.

- Allereerst worden scores van klein naar groot op volgorde gezet. Ook wordt genoteerd met wel percentiel elke waarde samengaat.
- Vervolgens moeten de z-waarden gevonden worden die met deze percentielen samengaan. Dit worden ook wel *z-normaal-scores* genoemd.
- Tot slot moet elke datapunt grafisch verbonden worden aan de corresponderende normaalscore. Als de distributie (bijna) normaalverdeeld is, dan zullen de datapunten bijna op een rechte lijn liggen. Systematische afwijkingen van de rechte lijn duiden op een niet-normaalverdeelde distributie. Uitbijters zijn datapunten die ver van het algemene patroon in de plot liggen.