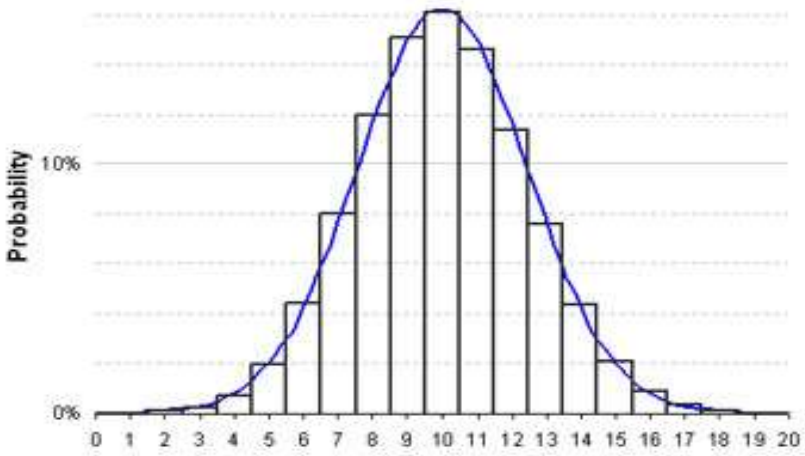
Hoofdstuk 5

De standaardnormale verdeling en diens eigenschappen

1. Inleiding

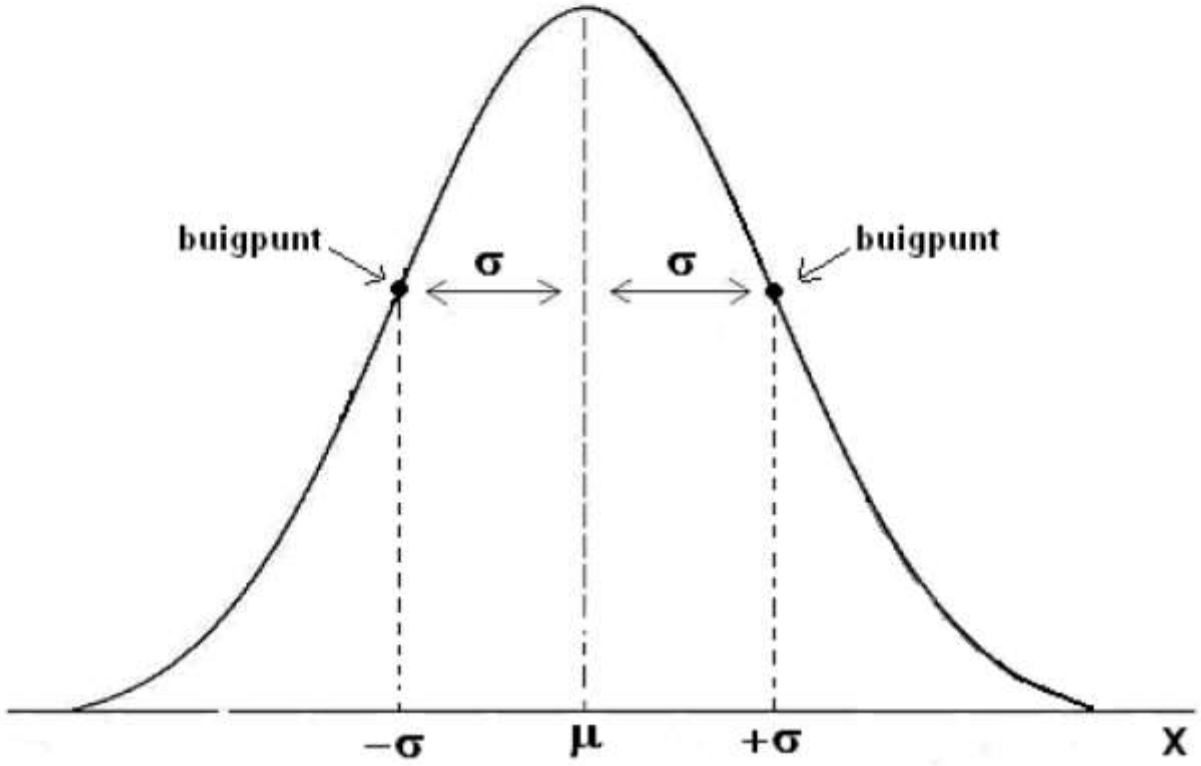
Wanneer we een continu gemeten kenmerk met haar verschillende waarden op de X-as plaatsen en op de Y-as de frequentie noteren, dan kunnen we blokjes tekenen die de frequentie weergeven. In de statistiek is men tot de vaststelling gekomen dat de aldus getekende histogrammen voor heel wat metrische kenmerken steeds een vergelijkbaar patroon hebben. Men vindt een beperkt aantal eenheden in de uiterste categorieën; de meeste waarnemingseenheden zullen eerder geconcentreerd rond het rekenkundige gemiddelde zitten. Deze verdeling noemt men ook wel **Gauss-curve of normale verdeling**. De verdeling benadert wat men noemt een **perfecte klokvorm**. Deze (theoretische) verdeling is uitermate belangrijk omdat er een aantal eigenschappen aan verbonden zijn, die we nodig hebben wanneer we later pogen om uitspraken op basis van steekproeven te veralgemenen naar de populatie. Elke proportie van een kenmerk komt overeen met wat we **een kans** (probabiliteit) noemen. De oppervlakte onder de curve stelt de proportie 100% voor.

Figuur: histogram met benadering van een normale verdeling



2. De normale en standaardnormale verdeling

Figuur: een normale verdeling van het kenmerk "gewicht", N (75,4)



Wat zien we in deze grafiek?

Vooreerst is het belangrijk stil te staan bij de notatie die we hier gebruiken. De notatie N (75,4) wijst op een verdeling met rekenkundig gemiddelde 75 en standaardafwijking 4.

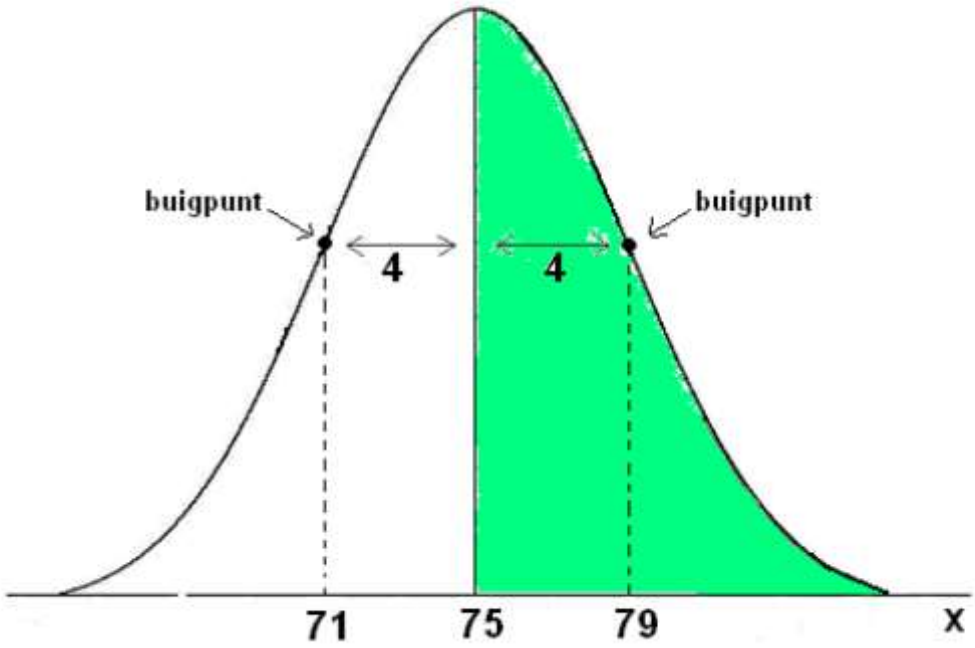
- De normale verdeling is gecentreerd om parameter μ of "mu", genoemd naar de Griekse letter. Deze parameter geeft de symmetrieas van de grafiek aan. Dit is het rekenkundig gemiddelde in een onderzoekspopulatie.
- De grafiek heeft twee ***buigpunten***, dat zijn punten waar de vorm van de kromme overgaat van 'bol' naar 'hol'. De tweede parameter van de normale verdeling σ of "sigma" bepaalt de ligging van deze buigpunten.

Uit wat hierboven werd gezegd trekken we een belangrijke conclusie: als we van een normale verdeling de twee parameters μ en σ kennen, kunnen we de grafiek van die normale verdeling tekenen. De hoogte van de grafiek is niet van belang zoals we later zullen zien.

In de normale verdeling (en in iedere continue verdeling) kunnen we een frequentie interpreteren als een oppervlakte onder de grafiek van de verdeling. Deze frequenties zijn te

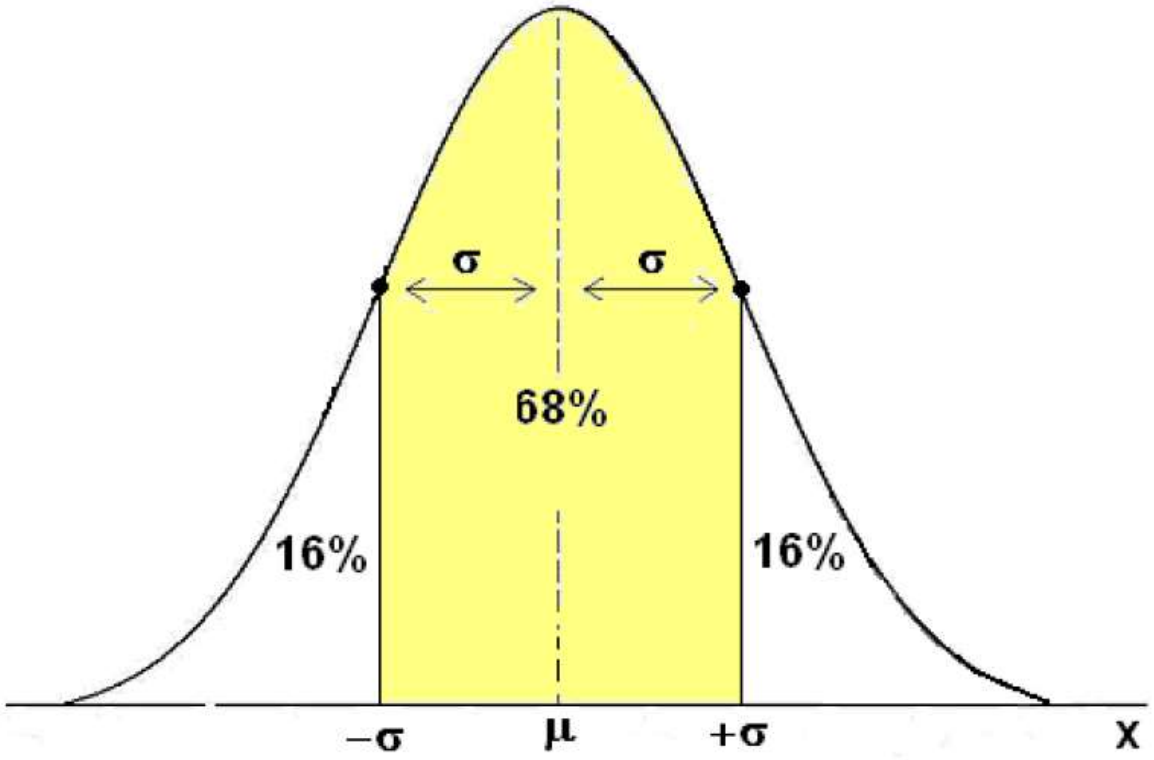
interpreteren als kansen, waarover specifiek meer in een volgend hoofdstuk. In een normale verdeling ligt 50% van de waarnemingen rechts van parameter μ, wat eenvoudig is te begrijpen omdat de normale verdeling symmetrisch rond μ is.

Stel bijvoorbeeld dat we weten dat in een populatie van 20-jarige Nederlandse jongens het gewicht normaal verdeeld is met een gemiddelde waarde van 75kg en een standaardafwijking van 4kg. De gewichten zijn hier op te vatten als numerieke waarden van de variabele 'gewicht', dus als scores. De vraag hoe groot het percentage jongeren is met een gewicht van meer dan 75kg is eenvoudig te beantwoorden. We tekenen eerst de grafiek van N(75,4) en geven in de grafiek het oppervlak rechts van μ = 75 aan. Rechts van het gemiddelde μ = 75kg ligt 50% van het oppervlak onder de grafiek. Dus is 50% van de jongeren zwaarder dan 75kg. Als we willekeurig ("at random") één jongere kiezen uit onze populatie is de kans dat hij/zij zwaarder is dan 75kg gelijk aan 0,5.

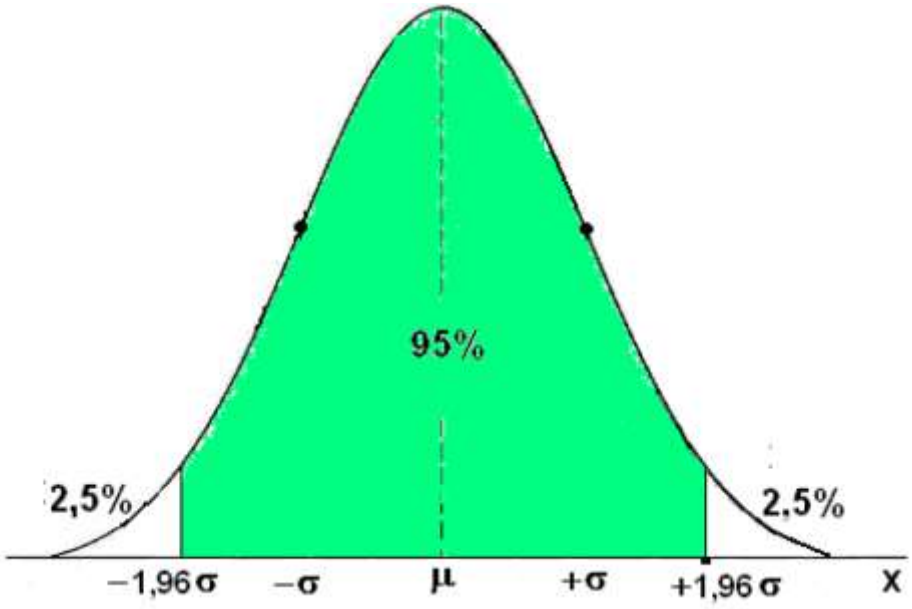


Om een schatting te kunnen maken van de proporties (= kansen) die overeenkomen met bepaalde waarden bij normaal verdeelde scores zijn de volgende drie eigenschappen van de normale verdeling van belang:

1. Bij een Normale Verdeling ligt ongeveer 68% van de scores tussen de benedengrens μ – 1σ en de bovengrens μ + 1σ , dus in het interval [μ – 1 σ ; μ + 1 σ].



2. Bij een Normale Verdeling ligt 95% van de scores in het interval [μ – 1,96 σ ; μ + 1,96 σ].



3. Bij een Normale Verdeling ligt 99,7% van de scores in het interval [μ – 3 σ ; μ + 3 σ].
De twee “staarten” bevatten dan elk 0,15% van de scores. De resolutie van het gebruikte

tekenprogramma is te grof om dit in een figuur aan te geven. Het is duidelijk dat het zéér uitzonderlijk is dat een normaal verdeelde score meer dan 3 standaardafwijkingen verwijderd is van het gemiddelde. Dit wordt ook de **'68-95-99' regel** genoemd. Deze regel is cruciaal.

3. Van normale verdeling naar standaardnormale verdeling

Elke normale verdeling is volledig symmetrisch en *unimodaal*, zodat gemiddelde, modus en mediaan samenvallen. Hieruit volgt dat 50 procent van de waarden boven, en 50 procent van de waarden onder het gemiddelde ligt. Voor elke normale verdeling geldt dat een vast percentage van de waarden ligt tussen het gemiddelde en een bepaald getal. Echter, **normale verdelingen verschillen in termen van gemiddelde en standaardafwijking**. Om te vermijden dat men telkens de oppervlakte onder de curve voor elke verdeling moet berekenen op basis van de formule, hebben statistici gezocht naar een **standaardisering**. Hierdoor diende men de oppervlakte eenmalig te berekenen en vast te leggen in een tabel. Door te standaardiseren wordt het gemiddelde gelijkgesteld aan 0 en de standaardafwijking aan 1. Deze berekeningen houden een verandering van schaal in. De standaardnormale verdeling (ook wel ***z-verdeling*** genoemd) is een bijzonder geval van de normale verdeling. De standaardnormale verdeling neutraliseert de zuiver numerieke verschillen tussen normale verdelingen en geeft een algemeen overzicht van de kansverdeling, onafhankelijk van de grootte van de waarden.

De standaardnormale **z-verdeling** wordt samenvattend als volgt gekarakteriseerd: (a) het gemiddelde van de standaardnormale verdeling wordt op nul gesteld door van iedere waarde het gemiddelde van de oorspronkelijke reeks af te trekken en (b) de standaardafwijking wordt op 1 gesteld door de absolute waarde van het bij (a) berekende verschil te delen door de standaardafwijking van de oorspronkelijke reeks. Dit wordt aangetoond in de volgende paragraaf.

4. Z-scores en het gebruik van de tabel van de standaardnormale verdeling

We kunnen iedere normale verdeling N(μ, σ) transformeren tot de zogenaamde 'standaardnormale verdeling', symbolisch weergegeven als 'N(0,1)'. Om dit te doen moeten we iedere x-score omzetten in een z-score. Hoe doen we dat nu? Een voorbeeld brengt verduidelijking. We onderzoeken de lengte van een aantal mensen. We weten dat de gemiddelde lengte 168cm is en de standaardafwijking 12cm bedraagt. We willen weten hoe groot de proportie (of het percentage) is van **de onderzoekseenheden die kleiner of gelijk**

aan 143 cm lang zijn (en dus in het hieronder gearceerde gedeelte vallen). Een z-score wordt als volgt berekend:

$$Z = \frac{X - \mu}{\sigma}$$

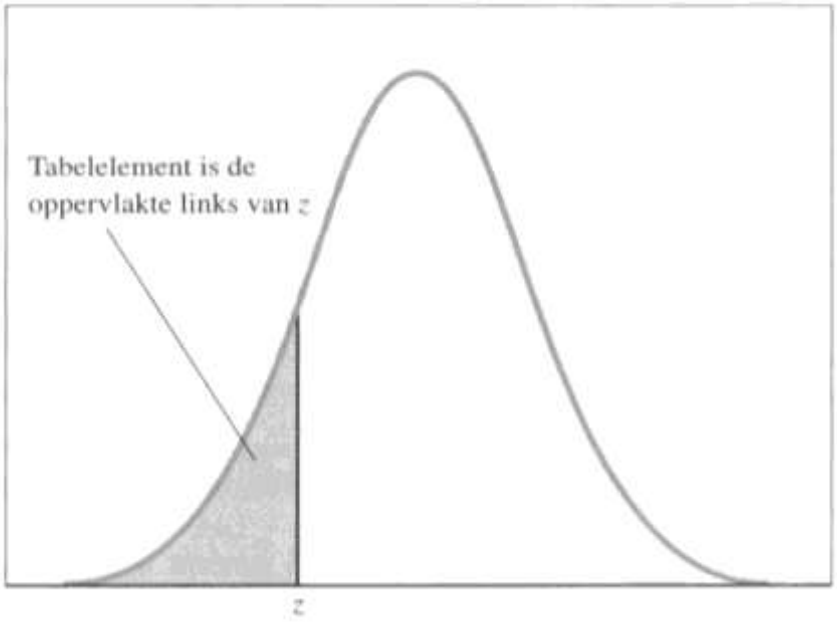
Voorbeeld 1

De waarde van Z voor iemand die 143 cm lang is, wordt als volgt berekend:

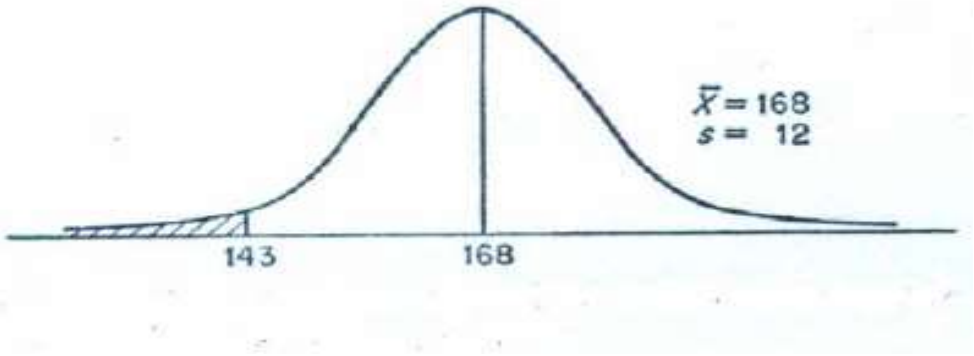
$$Z = \frac{143 - 168}{12} = \frac{-25}{12} = -2{,}08$$

Wat betekent nu deze waarde van -2.08? De oorspronkelijke ongestandaardiseerde waarde 143 ligt iets verder dan 2 standaardafwijkingen van het gemiddelde. Het negatieve teken van de Z-waarde wijst op het feit dat het gearceerde stuk links van het gemiddelde ligt. We kunnen in principe elke waarde van een metrische variabele standaardiseren. Dit betekent dat we voor elke onderzoekseenheden z-scores berekenen op de variabelen. **Voor het aflezen van de proportie van waarnemingen die in een bepaalde zone onder de normale valt, moeten we eerst de proportie van waarnemingen berekenen die overeenkomt met een score van 2.08. Dat is 98.12%. Aangezien de normale curve perfect symmetrisch is, weten we dat de proportie die overeenkomt met een z-score van -2.08 gelijk is aan 100% - 98.12% en dat is 1.88 %.** (Uit de tabel in bijlage kunnen de z-scores worden opgezocht).

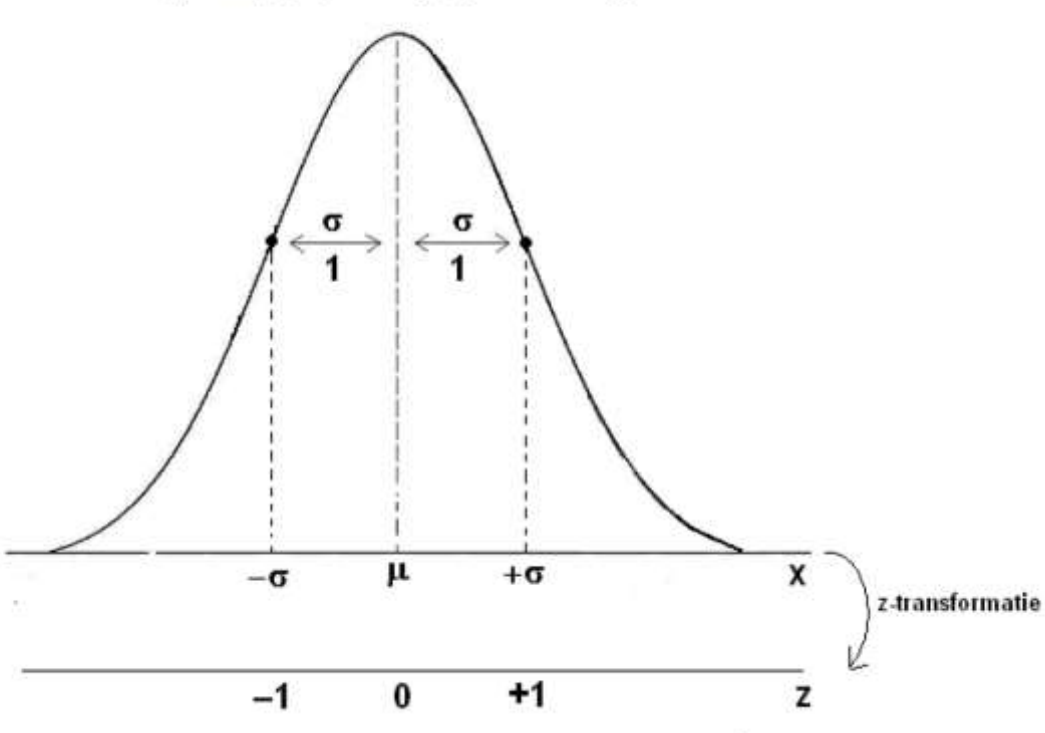
Figuur: opzoeken van tabelelementen



Figuur: voorbeeld bij het bepalen van een z-score

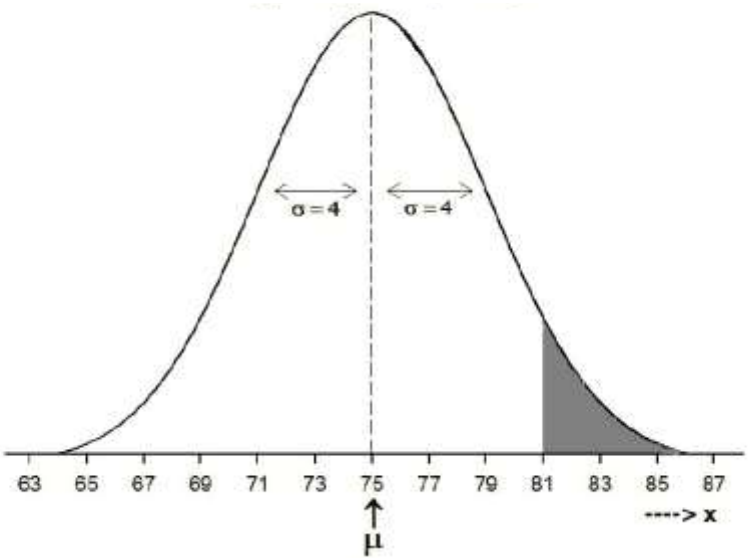


Samenvattend: Als de scores x normaal verdeeld zijn als N(μ , σ) dan zijn de hieruit berekende z-scores normaal verdeeld als N(0,1). Deze bijzondere normale verdeling noemen we de standaardnormale verdeling. Grafisch kunnen we het zo voorstellen dat je iedere normale verdeling kunt omzetten in een standaardnormale verdeling door de horizontale as te verschuiven ("centreren rond nul") en vervolgens uit te rekken of in te krimpen (standaarddeviatie van één). De volgende figuur geeft dit aan.



Voorbeeld 2

In een populatie is de variabele ‘gewicht’ normaal verdeeld met gemiddelde 75kg en standaardafwijking 4kg (symbolisch: x ~ N(75,4)). We trekken willekeurig (‘at random’) één persoon uit deze populatie. Wat is de kans dat deze persoon meer dan 81 kg weegt? Als we deze kans opvatten als een frequentie is de vraag: hoe groot is het oppervlak onder de grafiek van N(75,4) rechts van x = 81?



Z = (81 – 75) / 4 = 1.5. Als we deze kans opzoeken in de tabel van de standaardnormale verdeling, moeten we eerst zoeken welke waarde overeenkomt met 1.5 en dat is 0.9332. De kans om in deze populatie bij toeval iemand te trekken die lichter is dan 81 kg is dus 93.32%. De kans dat iemand zwaarder is dan 81 kg is dus 1 – 0.9332 of 0.0668 of 6.68%.

Voorbeeld 3

Gegeven een normaalverdeelde zwangerschapsduur van gemiddeld 280 dagen met een standaardafwijking van 10 dagen. Of anders genoteerd: N (280,10). Hoe groot is de kans dat de geboorte meer dan veertien dagen te laat plaatsvindt? We berekenen eerst de z-score die bij de vraag hoort: (294-280)\10= 1,4

Bij z = 1,4 hoort een kans van 91.92 %. Dit is de kans op een geboorte voor dag 294. De kans op een geboorte die méér dan 14 dagen te laat plaatsvindt, is dus 100 - 91.92 = 8.08 %

Voorbeeld 4

We vertrekken opnieuw van een normaal verdeelde zwangerschapsduur van gemiddeld 280 dagen met een standaardafwijking van 10 dagen. Hoe groot is de kans dat de geboorte plaatsvindt in de periode die ligt tussen 8 dagen voor en 8 dagen na de gemiddelde datum?

Dit voorbeeld lijkt moeilijker, maar dat is het eigenlijk niet. We moeten eerst weten welke de proportie is die overeenkomt met 288 dagen. Dit is 8 dagen na het gemiddelde. De z-score die hiermee overeenkomt is 0.8 want (288-280)/10 = 0.8. De proportie die hiermee overeenkomt is 0.7881. Vervolgens moeten we weten welke de proportie is die overeenkomt met 272 dagen. Dit is 8 dagen voor het gemiddelde. De z-score die hiermee overeenkomt is -0.8, want (272-280)/10 = -0.8. De proportie die hiermee overeenkomt is 1 – 0.7881 = 0.2119. De vraag behelst eigenlijk de proportie die tussen beide ligt: dus deze proportie bedraagt 0.7881-0.2119 en dat is 0.5762 of 57.62% wordt geboren tussen beide tijdstippen.

De normale verdeling (en andere verdelingen zoals de *binomiale verdeling* en de *chi-kwadraat verdeling* die we niet in detail bespreken in dit hoofdstuk) zijn belangrijk wanneer we uitspraken gebaseerd op één steekproef als onderdeel van een onderzoekspopulatie willen veralgemenen naar de totale onderzoekspopulatie. Hiertoe wordt gebruik gemaakt van de principes van de kansrekening en de inferentiële statistiek, die in volgende hoofdstukken worden uiteengezet. Kenmerken van statistische verdelingen worden dus pas echt belangrijk wanneer schattingen worden gemaakt en wanneer specifieke theorieën worden getoetst.

5. Leerdoelen

Dit deel beoogt studenten de centrale eigenschappen van normale verdelingen bij te brengen. Studenten dienen inzicht te hebben in deze eigenschappen en deze te kunnen toepassen. Deze toepassing houdt in dat men in staat is om vraagstukken op te lossen die gerelateerd zijn aan de berekening van oppervlaktes onder de curve van de normale verdeling. Het is belangrijk in te zien dat de normale verdeling een kansverdeling is die zegt hoeveel procent van de waarnemingen een bepaalde score hebben op een bepaald kenmerk. Het principe van de transformatie dient actief te kunnen worden toegepast. Dit hoofdstuk is belangrijk in functie van het begrijpen van de principes van de inferentiële statistiek, die in latere delen aan bod komen.