4. Kansverdelingen

Eerder is al besproken dat willekeur erg belangrijk is bij het verzamelen van data. Hierbij is vrijwel altijd bekend welke observaties mogelijk zijn, maar men weet nog niet welke er daadwerkelijk voor gaan komen. Kansen spelen daar een rol. De kans (*probability*) is de proportie van het aantal keren dat een bepaalde observatie voorkomt in een lange sequentie van observaties. De lange sequentie is hierbij belangrijk: naarmate deze langer wordt, wordt de kans steeds nauwkeuriger. De proportie uit je steekproef gaat dan steeds meer lijken op de proportie uit de populatie. Kansen kunnen ook worden weergegeven in percentages in plaats van in proporties.

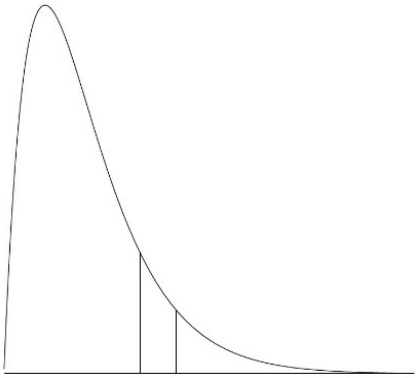
Kansformules
Een kans schrijf je vaak zo op: P(A). Hierbij is P de kans op uitkomst A. Stel dat er twee mogelijke uitkomsten zijn: A (getrouwd) en B (niet getrouwd). Dan schrijf je de kans op A als P(A). De kans op B staat gelijk aan 1 – P(A).

Stel dat er meerdere vragen worden gesteld en je wilt weten hoeveel van de getrouwde mensen ook gelukkig zijn. Dan vermenigvuldig je de kans dat iemand is getrouwd (A) met de kans dat iemand gelukkig is (B). De formule ziet er als volgt uit: P(A én B) = P(A) * P(B als ook A).

Kansverdelingen bij discrete en continue variabelen
Een discrete variabele heeft vastgestelde mogelijke waarden. Een continue variabelen kent ontelbare mogelijke waarden. Omdat een kansverdeling de kansen weergeeft bij elke mogelijke waarde van een variabele, gebeurt dit op verschillende wijze voor discrete en continue variabelen.

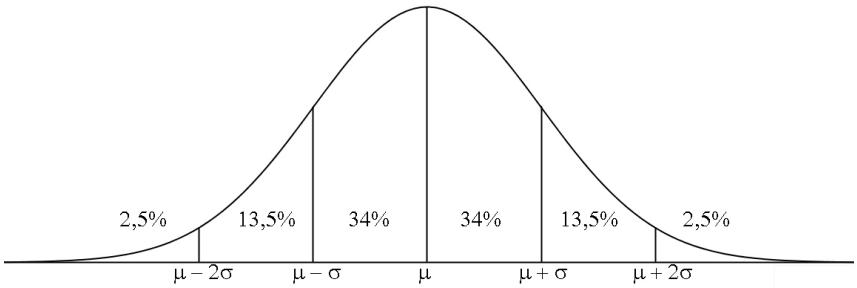
Bij een discrete variabele geeft de kansverdeling de kansen weer bij elke mogelijke waarde van de variabele. Elke kans is een getal tussen de 0 en de 1. De som van alle kansen staat gelijk aan 1. De kansen kunnen worden genoteerd als zijnde P(y). Hierbij is P de kans op een bepaalde waarde van y. In formule ziet dit er als volgt uit: $0 \leq P(y) \leq 1$, en $\Sigma_y P(y) = 1$.

Omdat een continue variabele een ontelbaar aantal mogelijke waarden heeft, kan de kansverdeling niet bij elke waarde een kans geven. Daarom geeft deze de kans weer van intervallen van mogelijke waarden. De kans dat een waarde binnen een interval ligt, ligt tussen de 0 en de 1. De kans dat alle mogelijke waarden binnen de interval liggen, is 1. Deze kansverdelingen worden weergegeven in een curve zoals hiernaast. Je ziet daarin een interval. Stel dat zich hierin 20% van de data bevindt, dan is de kans dat een waarde zich op die interval bevindt 0,20.



Net zoals bij een populatie distributie, heeft een kansverdeling parameters die de data beschrijven. Het gemiddelde beschrijft dan de centrale kans en de standaarddeviatie de variabiliteit. Het gemiddelde bij een discrete variabele kan worden berekend met de volgende formule: $\mu = \Sigma_y P(y)$. In woorden: je vermenigvuldigt alle mogelijke waarden met hun kansen, en deze tel je bij elkaar op. Deze parameter wordt ook wel de 'verwachte waarde van *y*' genoemd, en kan ook worden opgeschreven als E(y).

De normaal verdeling
De normaal verdeling is een type kansverdeling. Hij is erg belangrijk, omdat veel variabelen er in de werkelijkheid zo uit zien en omdat er heel veel statistische berekeningen mee gedaan kunnen worden. De normaal verdeling is symmetrisch, heeft een belvorm en heeft een gemiddelde (μ) en een standaarddeviatie (σ). De regels zijn nog steeds hetzelfde: 68% valt binnen 1 standaarddeviatie, 95% valt binnen 2 standaarddeviaties en 997% valt binnen 3 standaarddeviaties. De normaal verdeling ziet er dus zo uit:



Behalve dat er bij 1,2 en 3 standaarddeviaties vaste percentages horen, kan dat natuurlijk ook voor 1,5 of 1,7 of 1,9 (etc.) standaarddeviaties. Deze proporties en standaarddeviaties staan allemaal vast. Meestal geven we de hoeveelheid standaarddeviaties aan met *z*.

Voorbeeld:
Je hebt een variabele met μ = 18 en σ = 6. Je wilt weten hoe groot de proportie is die hoger heeft gescoord dan 30. De observatie is dus *y* = 30. Deze *y* moet je omzetten in een *z*-score. Dat doe je door (*y* - μ) / σ. In dit geval is *z* = (30 – 18) / 6 = 2. Nu kan worden opgezocht welke p-waarde er hoort bij een *z* = 2.

Steekproefverdeling
Bovenstaande gaat er van uit dat we weten hoe de populatie er uit ziet. In de werkelijkheid is dat meestal niet zo. We gebruiken daarom steekproeven. Met de statistieken uit de steekproeven kunnen we iets zeggen over de verwachte parameters uit de populaties.

Een steekproefverdeling geeft de kansverdeling van steekproefgrootheden (het is niet de verdeling van de uitkomsten in een steekproef). Elke statistiek heeft een steekproefverdeling (zoals een voor de mediaan, voor het gemiddelde, etc.). Het is een kansverdeling die de kansen weergeeft van de mogelijke uitkomsten van een statistiek. Een steekproefverdeling van een statistiek gebaseerd op *n* observaties is de relatieve frequentie verdeling van die statistiek, die het resultaat is van herhaalde steekproeftrekking van *n*, waarbij steeds de statistiekwaarde wordt berekend. Je kunt zo'n steekproefverdeling zelf maken door herhaalde steekproeftrekking, maar over het algemeen is de vorm van de verdeling wel bekend. Hiermee kun je dan de kansen van een waarde van een statistiek van een steekproef opzoeken bij een aantal (*n*) observaties.

Steekproefverdeling van het gemiddelde
Het gemiddelde is een veel gebruikte centrummaat. Maar wanneer het gemiddelde uit de steekproef bekend is, is nog niet bekend hoe dicht die ligt bij het gemiddelde van de populatie. Het is dus nog onbekend of $\bar{y} = \mu$. Maar omdat de steekproefverdelingen al bekend zijn, kunnen er toch uitspraken over gedaan worden. Bijvoorbeeld dat er een hoge kans is dat $\bar{y}$ binnen tien waarden van μ ligt. Echter, als er heel vaak een steekproef wordt getrokken dan zal blijken dat het gemiddelde van deze steekproeven gelijk is aan het gemiddelde van de populatie. Het gemiddelde van een steekproefverdeling is dan ook gelijk aan het gemiddelde van de populatie.

De spreiding van de steekproefverdeling van $\bar{y}$ wordt beschreven door de standaardfout van $\bar{y}$. Deze wordt genoteerd als $\sigma_{\bar{y}}$. De standaardfout kan worden berekend aan de hand van de volgende formule: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$. Voor een willekeurige steekproef met grootte *n*, hangt de standaardfout van $\bar{y}$ af van de standaarddeviatie van de populatie (σ). Uit de formule kan worden opgemaakt dat de standaardfout steeds kleiner wordt naarmate *n* groter wordt. Een grotere steekproef is dan ook een betere weergave van de populatie. Het feit dat het steekproefgemiddelde niet volledig overeenkomt met het populatiegemiddelde noemt men de steekproeffout. Deze wordt ook kleiner naarmate de steekproefgrootte (*n*) groter wordt.

Centrale limietstelling

Ongeacht de vorm van een populatiedistributie, de vorm van de steekproefverdeling van $\bar{y}$ is altijd een belvorm. Dit wordt de centrale limietstelling genoemd. Ook al is de populatiedistributie zeer scheef verdeeld, dan nog heeft de steekproefverdeling een belvorm. Wanneer de populatie echter zeer scheef verdeeld is, moet de steekproef wel steeds groter worden om meer deze belvorm aan te nemen. Hoe schever verdeeld de populatie, hoe groter de steekproef.