

2 Populaties, steekproeven, stochastische variabelen en kansverdeling

2.1 Populaties en de aselechte steekproef

In de statistiek is een populatie een ten aanzien van bepaalde aspecten homogene verzameling van objecten waarop het onderzoek zich richt². Als voorbeeld in dit hoofdstuk zullen in dit hoofdstuk schroeven worden genomen. De populatie bestaat dan uit het aantal schroeven dat een fabriek per kwartaal produceert. Ieder kwartaal zal de machine die de schroeven produceert opnieuw worden afgesteld om zo de gemiddelde, standaard lengte van iedere geproduceerde schroef te waarborgen. Als men de gemiddelde lengte van de schroeven wil onderzoeken gaat men uiteraard niet iedere schroef individueel onderzoeken. Dit zou immers veel te veel tijd kosten. In plaats daarvan doet men een steekproef en dit is tevens de term voor het aantal te onderzoeken schroeven. Belangrijk hierbij is dat deze steekproef aselechte is en dat houdt in dat iedere schroef die in een kwartaal wordt geproduceerd evenveel kans heeft om in de steekproef terecht te komen. Als gevolg zal iedere steekproef andere resultaten opleveren en door middel van de kansrekening kan men hier uitspraken over doen. (Let op! Uitspraken zijn geen conclusies!)

2.2 De stochastische variabele

De formele definitie van een stochastische variabele X is een (meetbare) reële functie in de uitkomstenruimte Ω ³. In het geval van de schroeven dat hierboven is beschreven zou men uit de uitkomstenruimte $\Omega = \{\text{alle in een kwartaal geproduceerde schroeven}\}$ de volgende stochastische variabele kunnen afleiden:

$X = \text{"Lengte, afgerond in millimeters, van een schroef"}$

Voor de stochastische variabele worden meestal hoofdletters gebruikt zoals X , Y en Z . Sommige statistische literatuur hanteert kleinere, onderstreepte letters zoals \underline{x} , \underline{y} en \underline{z} .

² http://nl.wikipedia.org/wiki/Populatie_%28statistiek%29

³ http://nl.wikipedia.org/wiki/Stochastische_variabele

2.3 Kansverdeling

Een kansverdeling is een tabel, formule of grafiek die de waarden van een stochastische variabele en de kans corresponderend met deze waarden beschrijft. Deze kansverdeling is van belang voor de statistische analyse die men wil uitvoeren en de kansverdeling wordt opgevat als een model voor de werkelijke verdeling, daar de werkelijke verdeling immers onbekend is. Het is de bedoeling van de statistische analyse om uitspraken te doen over de parameter(s) die de theoretische kansverdeling bepalen. Er bestaan verschillende vormen van uitspraken over parameters. Het kan (1) een schatting van een parameter betreffen, (2) gaan om het toetsen van een hypothese over een parameter of (3) het bepalen van een betrouwbaarheidsinterval voor de parameter.

2.4 Discrete en continue kansverdeling

Als een stochastische variabele discreet is, houdt dat in dat deze variabele een telbaar aantal uitkomsten kan hebben. Wanneer er bijvoorbeeld wordt gekeken hoe vaak men "kop" werpt als men 10 keer een euromunt opgooit, zijn de waarden van de stochastische variabele X 0, 1, 2, ..., 10. X kan in totaal 11 verschillende waarden aannemen.

De kansverdeling van een discrete stochastische variabele die waarden kan aannemen uit de verzameling K , wordt volledig bepaald door de kansen $p(k) = P(X = k)$.

Nu kunnen er twee vereisten voor discrete stochastische variabelen vastgesteld worden:

1. $0 \leq p(k_i) \leq 1$ voor alle k_i
2. $\sum_{k \in K} p(k) = 1$ (dit heet ook wel de kansfunctie)

Voorbeeld 2^a

Stephanie werkt bij een callcenter en heeft ingepland dat ze morgen 3 mensen gaat bellen. Ze weet uit ervaring dat de kans dat ze een klant binnenhaalt 20% per gesprek is. Geef de kansverdeling voor het aantal klanten dat morgen binnengehaald wordt. (Hint: maak eerst een kansboom)

Na het maken van een kansboom en het vaststellen van X = "aantal binnengehaalde klanten" kan men concluderen dat:

$$p(0) = 0,512$$

$$p(1) = 0,128 + 0,128 + 0,128 = 0,384$$

$$p(2) = 0,032 + 0,032 + 0,032 = 0,096$$

$$p(3) = 0,008$$

Als een stochastische variabele continu is, houdt dat in dat deze variabele niet telbaar is. Met andere woorden, de mogelijke waarden voor de variabele is oneindig. Een goed voorbeeld hiervan is de tijd die nodig is om een taak te voltooien. Bijvoorbeeld $X =$ "de tijd nodig om een tentamen te maken op de Universiteit waar 3 uur voor staat en studenten mogen pas na 1 uur weg". De kleinste waarde die X kan hebben is 60 (minuten). De grootste waarde voor X is 180. Tussen 60 en 180 kan X oneindig veel waarden hebben. Een student kan bijvoorbeeld na 60,1 minuten weg gaan, maar ook na 60,01 minuten. Of na 60,001 minuten of na 103,04245 minuten. Kortom, X heeft oneindig veel mogelijkheden en is daarom een continue stochastische variabele.

Indien meer dan 1 variabele een rol spelen, is er sprake van een zogenaamde stochastische vector. De stochastische vector is de combinatie van de stochastische variabelen, oftewel

$$\mathbf{X} = (X_1, \dots, X_k)$$