Populatieverdeling, verdeling van steekproefscores, en steekproevenverdeling

Als je data analyseert is het handig drie verdelingen te onderscheiden: de populatieverdeling, de verdeling van steekproefscores, en de steekproevenverdeling. Deze bron introduceert deze drie verdelingen.

De populatieverdeling

De populatieverdeling bevat de scores van alle leden in de populatie. Een voorbeeld van een populatie is bijvoorbeeld alle inwoners van Nederland. Het is belangrijk om er bij stil te staan dat zo'n populatie bijna altijd een theoretisch concept is: het is bijvoorbeeld niet mogelijk om op één moment alle leden van de populatie in een kamer te zetten. Dit komt omdat de dataverzameling en analyse van die data tijd kosten. Dit betekent dat als een studie start, het weken, maanden of soms zelfs jaren kan duren voordat er uitspraken gedaan kunnen worden over de verzamelde data. In die tijd kan de samenstelling van de populatie veranderd zijn: er verhuizen immers mensen naar Nederland, er verhuizen mensen uit Nederland, er worden mensen geboren, et cetera. Populaties zijn daarom meestal gedefinieerd als een groep mensen over tijd, bijvoorbeeld een spanne van een paar jaar. Hierdoor zijn uitkomsten van onderzoek nog steeds bruikbaar, ook als de data al een jaar oud zijn. Op de laatste pagina staat bovenin de verdeling van leeftijd voor alle mensen in Nederland in 2012 (bron: CBS). Deze verdeling heeft dus praktisch dezelfde vorm als de populatieverdeling, die gedefinieerd is als de verdeling van leeftijd van de inwoners van Nederland (even aangenomen dat er niet opeens meteorieten in Nederland inslaan).

De verdeling van steekproefscores

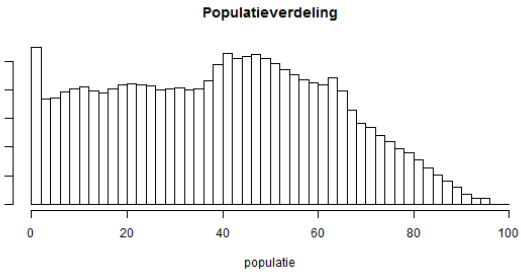
Omdat een populatieverdeling dus per definitie theoretisch is, is het onmogelijk om iedereen uit die populatie te meten. Desondanks willen we graag uitspraken doen over die populatie. De oplossing voor dit probleem is de aselecte steekproef: hierbij selecteren we willekeurig een aantal mensen uit de populatie. De laatste pagina laat in de linker kolom voorbeelden zien van steekproeven van 1, 2, 3, 4, 5, 10, en 50 mensen. Omdat die steekproef uit de populatie is getrokken, kunnen we op basis van die steekproef iets zeggen over de populatie. Een complicatie bij dergelijke uitspraken is dat de steekproef tot stand is gekomen op basis van toeval. Kijk bijvoorbeeld naar de steekproef van 1 persoon. Deze persoon is uitzonderlijk jong. We hadden ook toevallig een ouder persoon te pakken kunnen hebben. In onze steekproef van twee personen hebben we twee personen die toevallig in de buurt van het gemiddelde liggen; en in de steekproef van drie personen hebben we weer een uitzonderlijk jong persoon, maar ook twee mensen die juist iets boven het gemiddelde liggen. Gemiddeld genomen is deze steekproef best representatief: het gemiddelde ligt dicht in de buurt van het gemiddelde van de populatieverdeling. Je ziet dat naarmate de steekproef groter wordt, de verdeling van steekproefscores steeds dichter in de buurt van de

populatieverdeling komt. Met betrekking tot de schatting van het populatiegemiddelde en de standaarddeviatie in de populatie geldt dus dat de rol van toeval steeds kleiner wordt naarmate de steekproef groter wordt. Het zou mooi zijn als we ook informatie hadden over hoe accuraat de schatting van het populatiegemiddelde nu eigenlijk is. Gelukkig hebben we die informatie ook, in de vorm van de steekproevenverdeling.

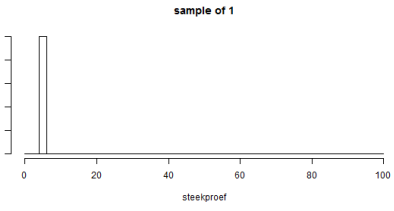
De steekproevenverdeling
Zoals je ziet in de verdeling van steekproefscores hebben die steekproeven elke keer een ander gemiddelde. Dit geldt natuurlijk ook als je dezelfde steekproef herhaalt: zie de vijf steekproeven van drie mensen in de middelste kolom op de laatste pagina. Zoals je ziet verschilt dat gemiddelde elke keer een beetje, maar ze liggen natuurlijk allemaal dichter in de buurt van het populatiegemiddelde dan wanneer we steekproeven van één persoon hadden genomen. Natuurlijk geldt het omgekeerde ook: als we steekproeven van tien mensen nemen, liggen de gemiddelden (gemiddeld genomen) nog dichter in de buurt van het steekproefgemiddelde. Er is natuurlijk nog wel een kans op extreme gemiddelden (we zouden heel toevallig tien keer een baby kunnen treffen), maar dit gebeurt natuurlijk zelden. We kunnen een beeld krijgen van hoe groot de kans op extreme gemiddelden nu eigenlijk is door de zogenaamde steekproevenverdeling op te stellen. Dit is een verdeling waar geen individuele scores in staan, maar juist de gemiddelden uit onze steekproeven – of beter gezegd, de gemiddelden van een oneindig aantal herhalingen van steekproeven met onze steekproefgrootte. Op de laatste pagina staan in de rechter kolom voorbeelden van die steekproevenverdelingen met verschillende steekproefgroottes.

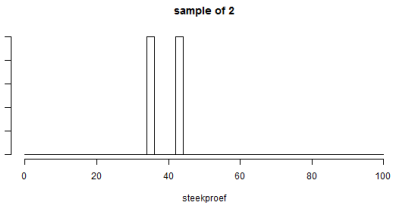
Je ziet hier dat de vorm van deze verdeling afhankelijk is van de steekproefgrootte. Als we een steekproef van één persoon nemen, en die oneindig vaak herhalen, is de steekproevenverdeling gelijk aan de populatieverdeling. Dat is natuurlijk logisch: het gemiddelde van een steekproef van één persoon is gelijk aan de score van die ene persoon. Het nemen van oneindig veel steekproeven van één persoon is dus hetzelfde als het nemen van een steekproef van oneindig veel personen, en dan krijg je natuurlijk weer de populatieverdeling. Naarmate het aantal mensen in onze steekproef toeneemt, krigjt de steekproevenverdeling steeds meer de vorm van een klok. Deze verdeling noemen we een normaalverdeling, en het is een verdeling die heel veel voorkomt. Hoe de populatieverdeling er ook uitziet, naarmate je een grotere steekproef neemt, gaat de steekproevenverdeling steeds meer op die normaalverdeling lijken: dit heet de *centrale limietstelling*.

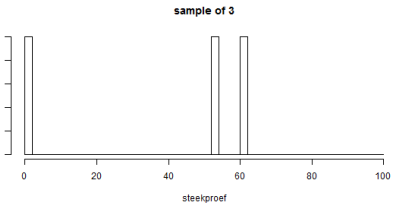
Het mooie is nu dat als je een steekproef neemt, het gemiddelde uit die steekproef eigenlijk uit die steekproevenverdeling komt. Die verdeling bevat immers alle mogelijke gemiddelden van een steekproef met een gegeven omvang. Als je die steekproevenverdeling kent, kun je dus uitrekenen hoe groot de kans is op een gemiddelde leeftijd die bijvoorbeeld 10 jaar onder de gemiddelde leeftijd in de populatie ligt: je weet immers welk percentage van de gemiddelde leeftijden in die steekproevenverdeling lager dan 10 jaar onder de gemiddelde leeftijd ligt. Deze kennis wordt dan weer gebruikt voor t-toetsen, en bovendien op een vergelijkbare manier voor andere toetsen, zoals F toetsen en χ^2 toetsen.

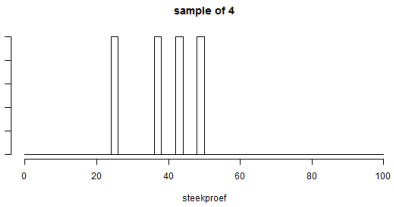


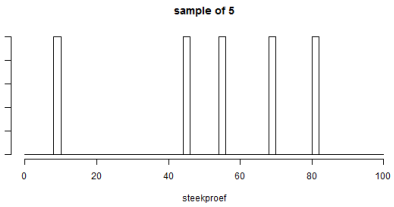
Steekproeven trekken

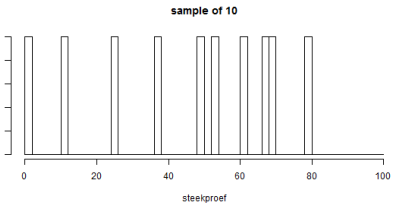


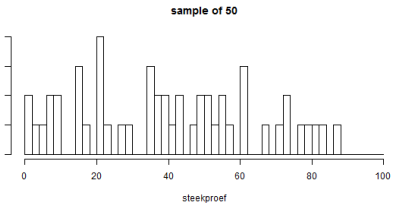


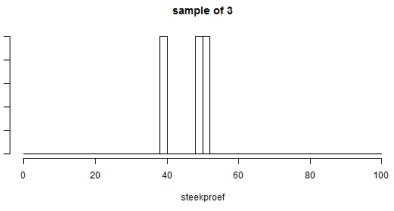


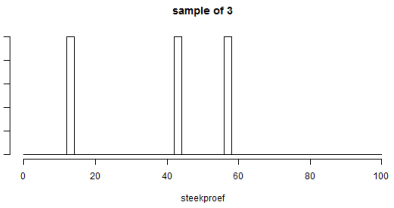


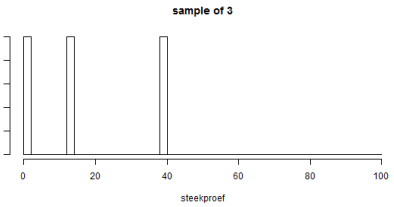


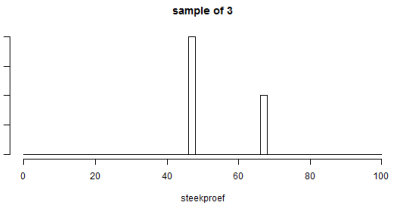


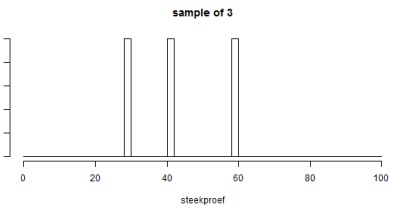




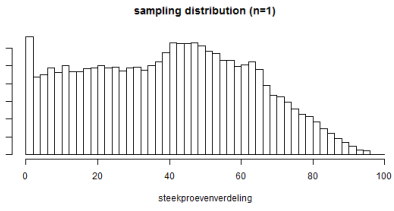


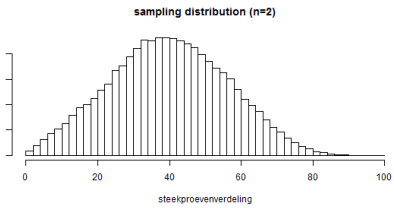


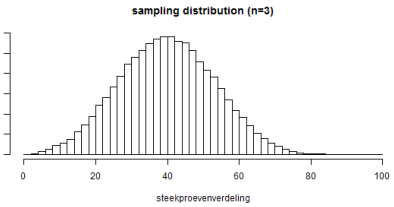


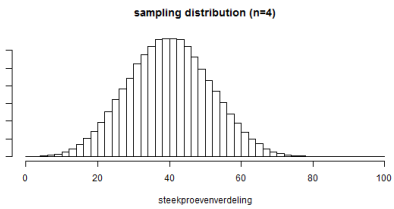


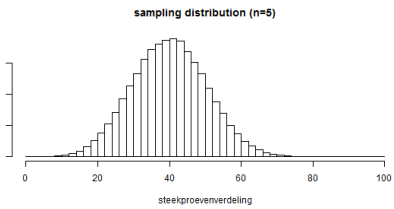
Vijf steekproeven van n=3

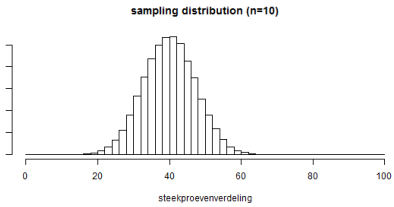


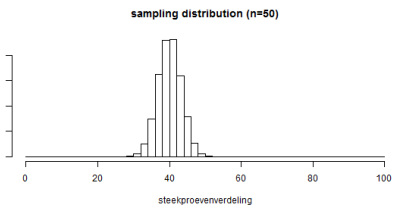












Steekproeven met verschillende n

Steekproevenverdelingen