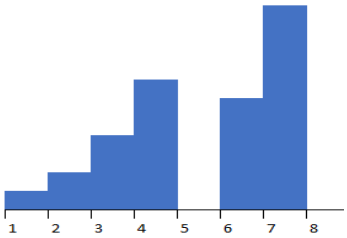
Data-analyse - “Een onrepresentatief gemiddelde … wat nu?”

De nieuwe huurdersbeoordelingen van 2019 zijn binnen en ieders instinct is om het gemiddelde te berekenen, zodat zichtbaar is hoe goed je als corporatie hebt gescoord. Alleen waarom bereken je het gemiddelde en wat zegt dit gemiddelde precies? Binnen de statistiek zijn er ook andere maatstaven zoals de mediaan en de modus. Alle drie zijn gevalideerde en bekende maatstaven maar wanneer is de ene techniek beter geschikt dan de andere?

In deze blog zal ik deze vragen beantwoorden door het uitleggen van een interessant en essentieel onderdeel binnen de statistiek genaamd *centrale tendentie*.

Verdeling

Voordat ik begin met het uitleggen van centrale tendentie en het belang daarvan, moeten we eerst wat weten over het begrip distributie dat ook wel verdeling wordt genoemd. Een distributie van je data is een lijst, tabel of een grafiek die de frequentie aantoont van de verschillende voorgekomen uitkomsten; dus hoe vaak heeft een bepaalde uitkomst zich voorgedaan. Waarom is de verdeling zo belangrijk? De verdeling helpt ons met het vinden van de centrale tendentie.

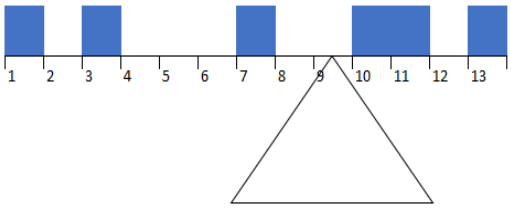


Figuur 1 Histogram

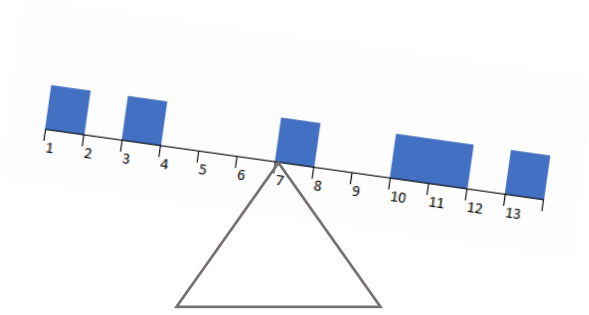
beoordeling	frequentie
1	1
2	2
3	4
4	7
6	5
7	10

Tabel 1 frequentietabel

De centrale tendentie is het meest centraal gelegen datapunt van de distributie. Een centraal punt kan worden beschouwd als het punt waarop de distributie in balans is. Als we de grafiek visualiseren als een weegschaal en de individuele datapunten zien als gewichten dan zoeken we naar een punt waarin de weegschaal perfect in balans is (zie figuur 2).



Figuur 2 centrale punt in balans



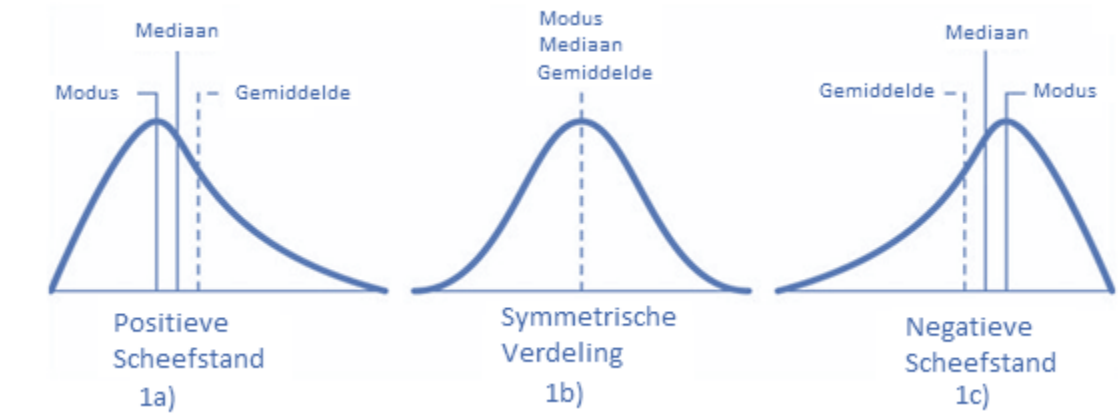
Figuur 3 Centrale punt in onbalans

Als we het centrale punt op 9,5 zetten dan is het mooi in balans, maar een verschuiving naar bijvoorbeeld rechts met een centraal punt van 7 brengt de weegschaal in onbalans (zie figuur 3). Wat betekent dat er is gekozen voor een centraal punt dat niet centraal gelegen is en dus niet de kern van

de dataset representeert. Het centrale punt kan op verschillende manieren worden bepaald maar dit is sterk afhankelijk van hoe je verdeling eruitziet.

Bij de meeste statistiche toetsingen wordt ervan uitgegaan dat er sprake is van een normale verdeling: een symmetrische verdeling (zie figuur 1b), met een gemiddelde genaamd μ als centraal punt. Voor een normale verdeling maakt het niet uit of er gebruik wordt gemaakt van een gemiddelde, mediaan of modus, omdat alle drie maatstaven met elkaar overeenkomen. Echter in de praktijk is er vaker sprake van een scheefstand in de verdeling: een dataset met meer spreiding en onregelmatigheden. Hierdoor kunnen de modus, mediaan en het gemiddelde van elkaar gaan verschillen.

Er is sprake van positieve scheefstand (Figuur 4a) wanneer de grootste hoeveelheid van de data zich aan de linkerzijde van de grafiek bevindt en negatieve scheefstand (Figuur 4c) wanneer de grootste hoeveelheid van de data zich aan de rechterzijde bevindt.



Figuur 4 voorbeelden verdeling

Nu je een beter begrip hebt over het belang van een verdeling gaan we kijken naar de drie verschillende maatstaven.

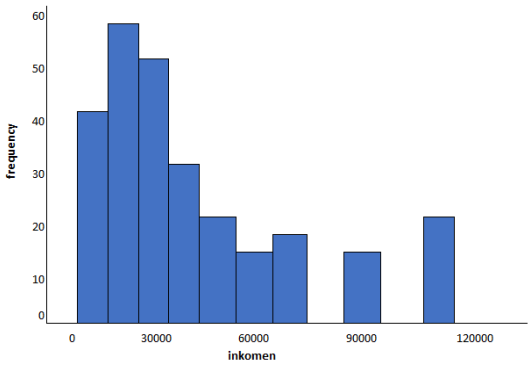
Gemiddelde

De eerste maatstaaf waar we mee beginnen is het rekenkundige gemiddelde. Dit is de som van alle waarnemingen gedeeld door het aantal waarnemingen:

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

Het kan gebruikt worden bij zowel continue als discrete data. Het neemt alle waarnemingen mee in de berekening en elke waarneming weegt even zwaar mee. Waardoor het gemiddelde de meest representatieve waarde is van de dataset.

Het nadeel van het gemiddelde is dat het gevoelig is voor uitschieters; dit zijn waarnemingen die minder frequent voorkomen en de spreiding van de data vergroot waardoor er een scheefstand ontstaat in de verdeling. Deze scheefstand heeft invloed op het gemiddelde en zorgt ervoor dat het minder centraal gelegen is.



Figuur 5 Histogram inkomensverdeling

In figuur 5 zie je een histogram van de inkomens van een aantal huishoudens. De verdeling toont een positieve scheefstand met een aantal uitschieters die de uiteindelijke spreiding van de data vergroot. Hierdoor is het gemiddelde rond 41.000 en dus veel minder centraal gelegen. Het is dus het beste om het gemiddelde te gebruiken wanneer de verdeling meer symmetrisch is gevormd.

Mediaan

De tweede meetstaaf is de mediaan, dit is de middelste waarde van alle waarnemingen:

$\frac{n+1}{2} = middelste\ positie$. Mocht het zo zijn dat er een even aantal waarnemingen zijn, dan neem je het gemiddelde van de twee middelste waarnemingen.

Hieronder in tabel 2 wordt het aantal van de huishoudgrootte getoond. De mediaan van dit voorbeeld is 2 en het rekenkundige gemiddelde van deze dataset is 2,22 Het gebruik van de mediaan of het gemiddelde zou beide kunnen.

Huishoudgrootte	**1**	**1**	**1**	**2**	**2**	**3**	**3**	**3**	**4**

Tabel 2 voorbeeld 1 Huishoudgrootte

Het voordeel van de mediaan is dat het een robuuste techniek is die minder vatbaar is voor uitschieters. Dit is vooral handig bij een kleine populatie, waarbij uitschieters een sterke invloed hebben. Wanneer er in het voorgaande voorbeeld twee huishoudgroottes aan worden toegevoegd met een aantal van twaalf en zestien (zie tabel 3) dan is de mediaan van 3 beter geschikt dan het gemiddelde van 4.4.

Huishoudgrootte	**1**	**1**	**1**	**2**	**2**	**3**	**3**	**3**	**4**	**12**	**16**

Tabel 3 voorbeeld 2 Huishoudgrootte

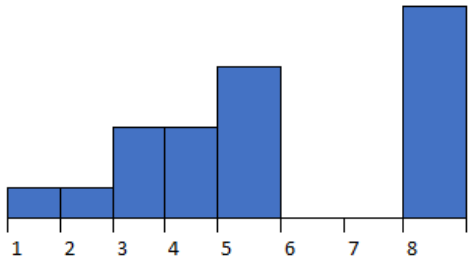
De mediaan is dus een hele handige manier om een centraal punt te bepalen dat niet al te veel wordt beïnvloed door uitschieters.

Modus

De derde meetstaaf om centraal gedrag van een dataset te tonen is de modus. Dit is de meest voorkomende waarde in een dataset en kan worden beschouwd als de populairste waarde. De modus wordt vaak gebruikt bij een categoriale variabele om zo te achterhalen wat de meest voorkomende categorie is.

Het nadeel van de modus is dat het mogelijk geen unieke waarde kent. Het zou kunnen dat twee waardes even frequent voorkomen en zelfs significant van elkaar verschillen. Het is dan lastig om aan

te duiden welke waarde gebruikt kan worden als centraal punt, echter de kans dat dit voorkomt bij een continue variabele is zeer klein.



Figuur 6 Histogram huurdersbeoordeling

Tenslotte kan het zo zijn dat de meest voorkomende waarde niet altijd centraal is gelegen. Figuur 6 toont een aantal beoordelingen (zie figuur 6). Zoals je kunt zien liggen de meeste beoordelingen onder de 6 maar het cijfer 8 is de meest voorkomende waarde. Het zou misleidend zijn om de modus te gebruiken om het centrale punt te bepalen.

Het is daarom aan te raden om de modus bij categoriale variabelen te gebruiken.

Wil je de centrale tendentie en andere statistiekonderwerpen toe kunnen passen in de praktijk? Schrijf je dan in voor onze nieuwe workshop data-analyse:
https://www.cns.nl/ondersteuning/workshop-data-analyse/