

5 Beschrijvende statistiek

5.1 Categorieën variabelen

In de statistiek onderscheidt men 4 hoofdgroepen variabelen: Nominaal, ordinaal, interval en ratio. De verschillen tussen deze 4 categorieën zijn redelijk eenvoudig aan te geven. Als men bijvoorbeeld kijkt naar een hardlooptwedstrijd met verschillende deelnemers, dan is een nummer van een renner een nominale variabele. Dit nummer heeft geen waarde en wordt alleen gebruikt voor het identificeren van individuele renners. Een ordinale variabele zou kunnen zijn de volgorde waarop de renners over de eindstreep binnenkomen, bijvoorbeeld als eerste, tweede, zesde, veertigste, enz. Hier heeft het getal weldegelijk een waarde, want het zegt iets over de volgorde van renners. Een interval variabele met betrekking tot de renners zou kunnen zijn dat de renners beoordeeld worden op een schaal van 0 tot 10, waar 0 heel slecht is en 10 heel goed. De ratio variabele kan bijvoorbeeld de tijd zijn waarin de renners de wedstrijd uitlopen en kan iedere waarde aannemen.

Naast categorieën variabelen heeft men ook een aantal maten nodig in de statistiek. Men onderscheidt respectievelijk locatiematen, spreidingsmaten en maten voor een lineaire samenhang.

5.2 Locatiematen

Locatiematen geven informatie over de locatie van het centrum van de verdeling. De belangrijkste locatiematen binnen de beschrijvende statistiek zijn modus, mediaan en gemiddelde. De modus wordt gedefinieerd als de waarde (of waarden) die het meeste voorkomt. De mediaan wordt berekend door alle waarden op volgorde te zetten (van laag naar hoog of van hoog naar laag). De waarde in het midden is de mediaan. Het gemiddelde wordt simpelweg berekend door de waarden bij elkaar op te tellen en te delen door het aantal waarden.

⁴ http://nl.wikipedia.org/wiki/Centrale_limietstelling

Voorbeeld 5a

Een steekproef bestaande uit 10 jongeren wordt gevraagd hoeveel uur per week ze computerspelletjes spelen. De resultaten staan in de tabel hieronder.

0	7	12	5	33	14	8	0	9	22
---	---	----	---	----	----	---	---	---	----

Het gemiddelde wordt als volgt berekend:

$$\frac{0 + 7 + 12 + 5 + 33 + 14 + 8 + 0 + 9 + 22}{10} = \frac{110}{10} = 11$$

De mediaan vindt men als men de waarden op volgorde zet van hoog naar laag:

0	0	5	7	8	9	12	14	22	33
---	---	---	---	---	---	----	----	----	----

De mediaan ligt dan tussen de 5^{de} en 6^{de} waarde, te weten 8 en 9, en daarom is de mediaan gelijk aan 8,5.

De modus is 0, want dit is de enige waarde die meer dan een keer voorkomt.

Het is van belang om te weten welke locatiematen van toepassing zijn op de verschillende categorieën variabelen. Op nominale variabelen is alleen de modus van toepassing. De modus en mediaan spelen een rol bij ordinale variabelen. Zowel modus, als mediaan, als gemiddelde zijn van toepassing op interval en ratio variabelen.

5.3 Spreidingsmaten

Deze maten hebben als doel om aan te geven in welke mate de waarden van een verdeling of steekproef uiteenlopen en zijn alleen van toepassing op kwantitatieve interval en ratio variabelen. De meest gebruikte spreidingsmaten zijn het bereik, de standaarddeviatie en de variatiecoëfficiënt. Het bereik is niets anders dan het verschil tussen de grootste waarde en de kleinste waarde. Soms is het bereik oneindig en dus onbruikbaar. De standaarddeviatie is al gegeven in hoofdstuk 3. De

variatiecoëfficiënt van een verzameling waarden is de standaarddeviatie van de waarden gedeeld door hun gemiddelde:

$$V = \frac{\sigma}{\mu}$$

5.4 Maten voor lineaire samenhang

In §3.3 zijn de begrippen covariantie en correlatiecoëfficiënt al besproken. In deze paragraaf worden de covariantie en correlatiecoëfficiënt met betrekking tot steekproeven behandeld (de zogenaamde steekproefcovariantie s_{xy} en steekproefcorrelatiecoëfficiënt r_{xy}).

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} (\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y})$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$