Hoorcollege: De normale verdeling; z-scores

Adolphe Quetelet:

Hij was degene die erachter kwam dat mensen eigenschappen hebben die, als je die registreert, een normale verdeling vormen. De Queteletindex is een andere naam voor de body-mass index. Hij had uitgevonden dat er een samenhang is tussen lengte en gewicht. Hij had hierdoor invloed op de statistiek met het concept van de (normale) natuurlijke variantie.

Dichtheidcurves:

Een histogram geeft exact de verdeling (in intervallen) weer, maar men kijkt liever naar een versimpelde verdeling. Daarom moet er gekeken worden of de verdeling is weer te geven als een gladde curve, ook wel dichtheidscurve of 'density curve' genoemd. Het is een vloeiende lijn door de toppen van het histogram. Curves hebben een algemene vorm.

De dichtheidscurve is een benadering van een empirische verdeling. De lijn geeft een goede benadering van het histogram weer.

Bij een normaal verdeling zijn de gemiddelde verdeling en de standaard deviatie genoeg om een model te kunnen tekenen. Een normaal verdeling is altijd symmetrisch en heeft één top. De normale verdeling is een model maar ook een norm.

Cumulatieve frequentieverdeling: hoe een verdeling aangroeit

Relatieve frequentieverdeling: proporties

Het totale oppervlak onder de curve is altijd 1. Als staafjes een relatieve frequentie voorstellen, kan je ze optellen en er iets over zeggen. Dit betekend dus ook dat ze in een histogram staan (week 4). Je kunt via 2 benaderingen een berekening doen naar cumulatieve staafjes: er zit wel een kleine onnauwkeurigheid in, maar deze is niet groot genoeg om je er zorgen over te maken.

Manier 1: de staafjes optellen, dus het oppervlakte van de staafjes bij elkaar optellen. Manier 2: het oppervlak onder de curve berekenen. De staafjes zitten soms wat boven de curve en soms wat eronder en dit reken je weg door alleen te kijken naar het oppervlak onder de curve.

Er bestaat wat abstractie boven wat je observeert. Daarom gebruiken we nu Griekse symbolen voor de standaardafwijking en het gemiddelde.

Eigenschappen normaalverdeling:

- Ze zijn klokvormig, ze hebben 1 top, ze zijn symmetrisch

- μ (het gemiddelde) is de locatie van de top, maar zegt verder niks over de vorm.

- σ is zichtbaar. Het is de afstand van de top (het gemiddelde) tot het buigpunt.

- Hoe wijder de standaardafwijking, hoe kleiner het gemiddelde. Dit komt omdat het oppervlak van de grafiek altijd 1 is en dus constant. Bij een grotere spreiding, krijg je dus een lagere top en andersom.

- μ doet verder niks, heeft geen invloed op het maxima, dat doet alleen σ. Dit geeft de vorm van de grafiek aan.

Hernstein& Murray:

Deze 2 onderzoekers deden samen bepaalde aannames:

Een hiervan is: IQ is normaal verdeeld en genetisch bepaald. De vraag is of dat zo is. We ontkennen veel dat diefstal enzovoorts genetisch bepaald is, omdat dit voor ons een angstaanjagende gedachte is. Deze aanname is aan te vallen met tweelingonderzoek. Hierbij kun je namelijk stellen dat tweelingen identiek zijn qua erfelijkheid en dus zouden ze hetzelfde moeten doen op het gebied van criminaliteit. Maar dit blijkt in de werkelijkheid niet het geval te zijn.

Toepassingen normaalverdeling:

Steekproefvariabiliteit: dit zijn de verschillen in steekproeven. Het zegt iets over de theoretische variabiliteit van de steekproef. De normaalverdeling is hier het model voor.

Standaardregel:

Dankzij de symmetrie en de standaardafwijking kan je in iedere normaalverdeling kijken naar bepaalde gebieden in de verdeling. Voor de vraag tussen welke x-waardes liggen hoeveel mensen kun je de 68-95-99.7% regel gebruiken.

Tussen 1 standaardafwijking links en 1 standaardafwijking rechts van het gemiddelde ligt 68% van de individuen onder de normaalverdeling; dus 68% van de individuen ligt tussen deze grenzen. Bij 2 standaardafwijkingen links en rechts is dat 95% en bij 3 standaardafwijkingen links en rechts is het 99.7%

Alle normaalverdelingen zijn in principe hetzelfde, omdat we er hetzelfde mee omgaan. Daarom kan er een standaard normaalverdeling worden gemaakt, waarmee je in iedere verdeling kan uitrekenen wat een bepaald stukje in de verdeling is.

Z-scores:

Z-scores zijn de standaard waarden die bij de standaard normaalverdeling horen. De z-score geeft aan hoeveel standaarddeviatie een score afwijkt van het gemiddelde. Z-scores heb je nodig bij het standaardiseren van data. Bij deze waarde op de x-as hoort een cumulatieve waarde. Deze scores zijn te vinden in de tabel in je werkboek.

Ruwe scores omzetten in z-scores:

$$z = \frac{x - \mu}{\sigma}$$

Van z-scores terug naar ruwe data (ontstandaardiseren):

$x = z * \sigma + \mu$.

Bij een standaard normaalverdeling is het gemiddelde altijd 0 en de standaarddeviatie altijd 1. Dus N(0,1). Is de z-score negatief, dan ligt deze links van het gemiddelde (gem. is 0). Positieve z-scores liggen rechts van het gemiddelde.

Belangrijk om te onthouden is dat standaardiseren alleen kan als de oorspronkelijke verdeling al een normaalverdeling was. Dit kan je snel zien door de ruwe data te plotten in een histogram en globaal te checken of het histogram ruwweg symmetrisch is en één top heeft (kan je er een density curve door tekenen?). Er veranderd dus niets aan de vorm, standaardiseren is alleen om scores met elkaar te kunnen vergelijken ten opzichte van het gemiddelde. Met standaardiseren kun je scores omzetten in kansen en andersom. Met standaardiseren bewerkstellig je dus geen normaalverdeling.

Tabel A:

Als je bij een Tabel A bijvoorbeeld aan de linkerkant naar beneden naar beneden zou gaan, zou je de hele cijfers van het getal zien, en wanneer je bovenaan in de tabel naar rechts gaat, kun je het getal aanvullen met eventuele decimalen.

Dus als je bijvoorbeeld bij een z-score van -0.22 een proportie wilt weten, ga je eerst naar -0.2 (dat doe je door naar beneden te gaan in de tabel). Daarna zoek je de decimalen hierbij, dus ga je naar 0.02 en kom je op een proportie van 0.4129.

Normal Quantile plot:

Deze plot is bedoeld om na te gaan of je data normaal verdeelt zijn. Je kunt dit nagaan door je data en de bijbehorende z-scores te plotten in een normal quantile plot. Wanneer deze plot een rechte lijn weergeeft, heb je te maken met een normaalverdeling. Je berekent de relatieve frequenties.