Het instrument

Aannemelijk maken van samenhang

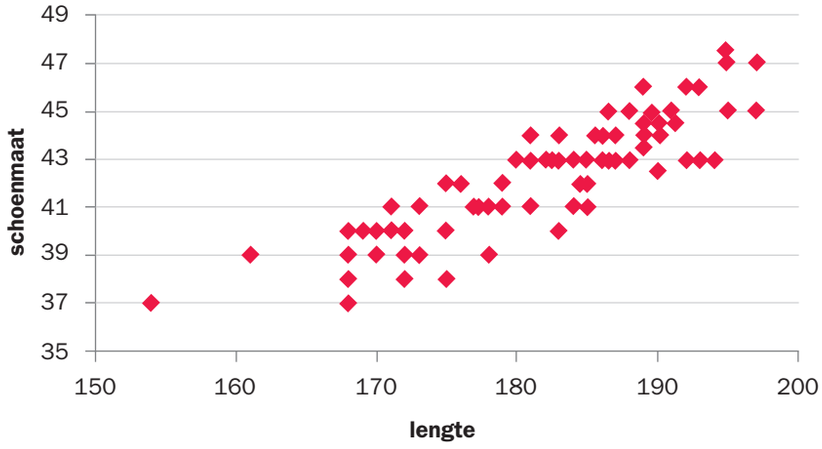
In de vorige bijdrage hebben we een inventarisatie gemaakt van de factoren die mogelijk van invloed zijn op het kenmerk dat we onderzoeken. Nu gaan we kijken of we dat ook aannemelijk kunnen maken op basis van data. Ik spreek bewust niet van 'bewijzen', want met statistische methoden doen we dat niet: er is altijd een kans dat het toch anders is.

Of kenmerken met elkaar samenhangen is statistisch wel aannemelijk te maken, maar je moet er ook een verklaring voor kunnen geven. Er moet een mechanisme zijn dat deze samenhang veroorzaakt. Zo is er een periode dat je salaris toeneemt met het aantal dienstjaren; het achterliggende mechanisme is het gebruik van de salarisschalen per dienstjaar. Er kan ook visueel een aanwijzing zijn, maar zonder verklaring is dat niet veel waard.

Als twee kenmerken met elkaar samenhangen, dan moet het zo zijn dat verandering van de waarde van het ene kenmerk invloed heeft op de waarde van het andere kenmerk. Dat is ook wat we zoeken als het gaat om vaststellen van oorzaak en gevolg uit het visgraatdiagram.
Om het netjes vorm te geven spreken we het volgende af. We onderzoeken de samenhang tussen twee kenmerken (variabelen), waarbij Y de variabele is die het gevolg is van de X variabele. Dus bijvoorbeeld als X (aantal dienstjaren) toeneemt, volgt ook een toename van Y (salaris). Het kan natuurlijk ook andersom, als X toeneemt kan Y afnemen (hopelijk niet bij je dienstjaren en salaris).

Ook de vorm van de samenhang is niet vanzelfsprekend. Sommige mensen hebben de 'trendlijn maken'-functie van Excel ontdekt en trekken door iedere puntenwolk een rechte lijn, of soms nog erger, een gekromde lijn. Zonder dat ze een verklaring geven van de werking van het achterliggende mechanisme. Die zou er wel kunnen zijn. Celgroei bijvoorbeeld is niet een lineair verschijnsel in tijd, dus daar zou een rechte lijn niet passend zijn, er ligt een ander groeimodel onder.

schoenmaat vs lengte



Grafische weergave

Om samenhang grafisch weer te geven is eigenlijk alleen maar een spreidingsdiagram van toepassing. Soms zie ik wel eens dat mensen in een grafiek een kolomdiagram tekenen van verschillende variabelen, 'omdat ze ongeveer wel dezelfde vorm vertonen', maar dat is niet de meest geschikte vorm.
In een spreidingsdiagram zetten we de oorzaak (variabele X) op de X-as en het gevolg (variabele Y) op de Y-as. Van iedere entiteit in onze dataset plotten we dan het puntenpaar (x,y) in de grafiek, net zoveel punten als we waarnemingen hebben.

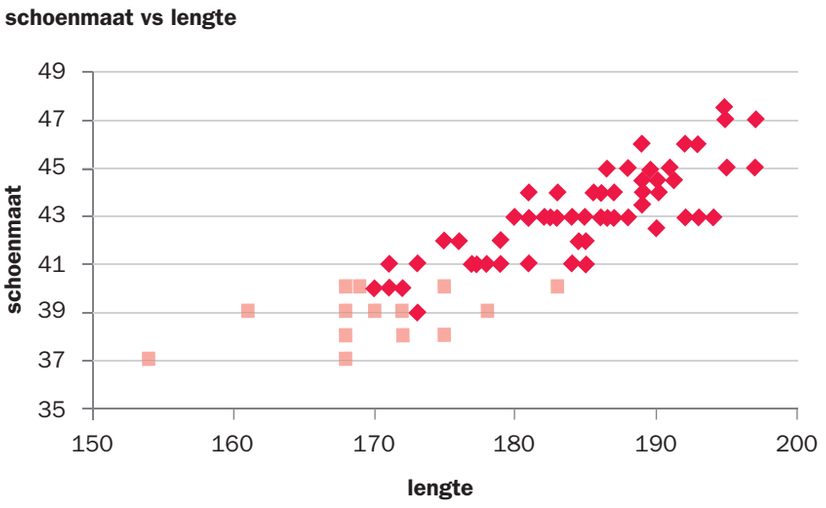
Voorbeeld

Langere mensen hebben waarschijnlijk ook een grotere schoenmaat. Om dat te onderzoeken vragen we verschillende mensen naar hun lengte en naar hun schoenmaat en zetten de resultaten in het genoemde spreidingsdiagram (figuur links). Het lijkt er dus wel op dat naarmate de lengte toeneemt, ook de schoenmaat toe neemt.

Naast deze simpele weergave is het ook mogelijk om bijvoorbeeld onderscheid te maken tussen mannen en vrouwen. In Excel geef je dan twee X-variabelen op, in de ene de data van de mannen en in de andere de data van de vrouwen. Automatisch wordt er dan in twee kleuren een puntenwolk weergegeven (figuur pag 25 bovenaan).

In versie Excel2013 is het ook mogelijk om de punten te voorzien van een individueel label dat in de grafiek zichtbaar wordt. Via Gegevenslabels -> Meer opties -> Labelopties -> Waarde uit cellen geef je de datareeks op waar de labels staan van deze waarnemin-

Arend Oosterhoorn is al vele jaren actief in de wereld van kwaliteitsmanagement en Lean Six Sigma. Hij begeleidt organisaties die op zoek zijn naar verbetermogelijkheden.
aoosterhoorn@oosterhoornadvies.nl

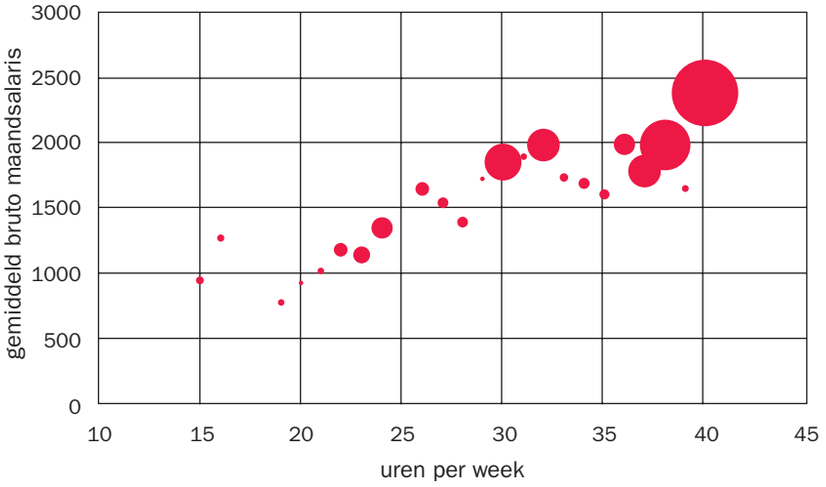


gen, deze worden dan bij ieder punt weergegeven. Met name bij afwijkende waarnemingen in de grafiek is dat makkelijk, je kunt direct terugzien welke waarnemingen het betreft.

Bellengrafiek

Het komt ook voor dat je van bepaalde combinaties meerdere waarnemingen hebt. Zou je de grafische weergave doen als boven, dan zie je deze meerdere waarnemingen slechts als één punt in de grafiek terug en dat is uiteraard ongewenst.

Dit kun je ondervangen door een bellengrafiek te maken. Dat is ook een spreidingsdiagram, maar de punten worden dan 'opgeblazen', al naar gelang er meerdere waarnemingen zijn van bepaalde combinaties (figuur hieronder). Dit werkt bijvoorbeeld erg goed als je de gemiddelden van de ene variabele (bijvoorbeeld bruto maandsalaris) uit wilt zetten tegen de waarde van een andere (bijvoorbeeld aantal werkuren per week). Het aantal waarnemingen in iedere klasse van werkuren is dan de 'opblaasfactor'.



Kengetal

Samenhang is niet alleen grafisch weer te geven, het is ook samen te vatten in een kengetal, de zogenoemde correlatiecoëfficiënt. Ik zal u niet bezwaren met formules, maar slechts verwijzen naar de berekeningswijze in Excel. De correlatiecoëfficiënt is een getal tussen -1 en + 1 en in de buurt van 0 is er geen sprake van samenhang. Naarmate de correlatiecoëfficiënt dichter bij -1 of +1 komt, is de samenhang sterker. Als de correlatiecoëfficiënt kleiner dan nul is, spreken we van een negatieve correlatie (de waarde van Y neemt af als de waarde van X toeneemt), is deze groter dan nul dan duiden we dat aan als positieve correlatie (de waarde van Y neemt toe met toenemende waarde van X). Er bestaan tabellen die aangeven wat de ondergrens is voor de waarde voordat je mag spreken van echte samenhang (uiteraard in combinatie met de verklaring). Ter illustratie, de samenhang tussen schoenmaat en lengte voor alle personen gezamenlijk komt uit op 0,82.

Vervolg

Onderzoek naar samenhang is essentieel in het verbeteren van de kwaliteit van producten/diensten en processen. Je moet tenslotte begrijpen hoe de mechanismen werken om het proces te kunnen beïnvloeden en bij te sturen. Er zijn ook verschillende vormen van een dergelijk onderzoek naar samenhang. Heb je daarover vragen, laat maar weten.

Voorlopig verlaten we de lijn van oorzaak en gevolg. Volgende keer kijken we naar het in kaart brengen van structureren en processen.