Dooreenschudden van gegevens om de signifikantie van een verband te schatten

Koen Van de moortel, 2008 04 05

Stel, ik vermoed het bestaan van een verband tussen grootheid x en grootheid y: y=f(x).
Ik beschik over een reeks meetgegevens (x_i, y_i). Aangezien het meetgegevens zijn, met dus meetfouten, zal het waarschijnlijk zelden voorkomen dat $y_i=f(x_i)$, maar zal $y_i \approx f(x_i)$.
Als we iets weten over het waarom van het verband tussen x en y, weten we normaal ook iets over de vorm van f, bv. lineair (y=ax+b, 2 onbekende parameters), exponentieel (y=a*exp(b(x-c))+d, 4 parameters), sinusoïdaal (y=a*sin(b(x-c))+d, 4 parameters),...

Het beantwoorden van de vraag “Welk verband van hogergenoemde vorm bestaat er bij mijn gegevens?” begint dan met het aanpassen van parameters van f, zodanig dat de meetgegevens zo dicht mogelijk bij de grafiek van f komen te liggen (kleinste kwadratenregressie).
Bij een dergelijke regressie wordt ook een getal berekend dat aangeeft hoe goed de gevonden f aansluit bij de gegevens. Bij lineaire regressie is dat bv. de zgn. “korrelatiekoëfficient” die dicht bij 1 ligt als de meetpunten inderdaad min of meer op een rechte lijn liggen, en dicht bij 0 als er helemaal geen lineair verband is. Algemener geeft de gemiddelde waarde van $(y_i-f(x_i))^2$ een goede schatting van de kwaliteit van de aanpassing (“goodness of fit”); ik zal deze hier verder s^2 noemen. Hoe kleiner s, hoe beter dus.

Als nu echter de hoofdvraag is of er eigenlijk wel een verband bestaat van de vorm y=f(x) bij mijn gegevens, dan is bijkomend onderzoek nodig.
Als f (juister: de klasse van funkties van een zekere vorm) inderdaad het verband weerspiegelt tussen mijn gegevens (en f is geen triviale konstante funktie), dan betekent dat dat mijn y-meetwaarden “in de juiste volgorde liggen”, d.w.z. dat de y-meetwaarden horen bij de x-meetwaarden, volgens het verband y=f(x). Als ik ook maar één koppel y-meetwaarden onderling zou verwisselen, zou het moeilijker worden om een dergelijke f aan de gegevens aan te passen, d.w.z.: s zou groter worden.
Een simpel voorbeeld (met het minimum van 1 vrijheidsgraad):

x_i	y_i
2005	100
2006	109
2007	120

Stel dat we hier een lineair verband vermoeden. Met lineaire regressie verkrijgen we hier een korrelatiekoëfficiënt (r^2) van 0.99667... wat doorgaans als “zeer goed” bestempeld wordt.

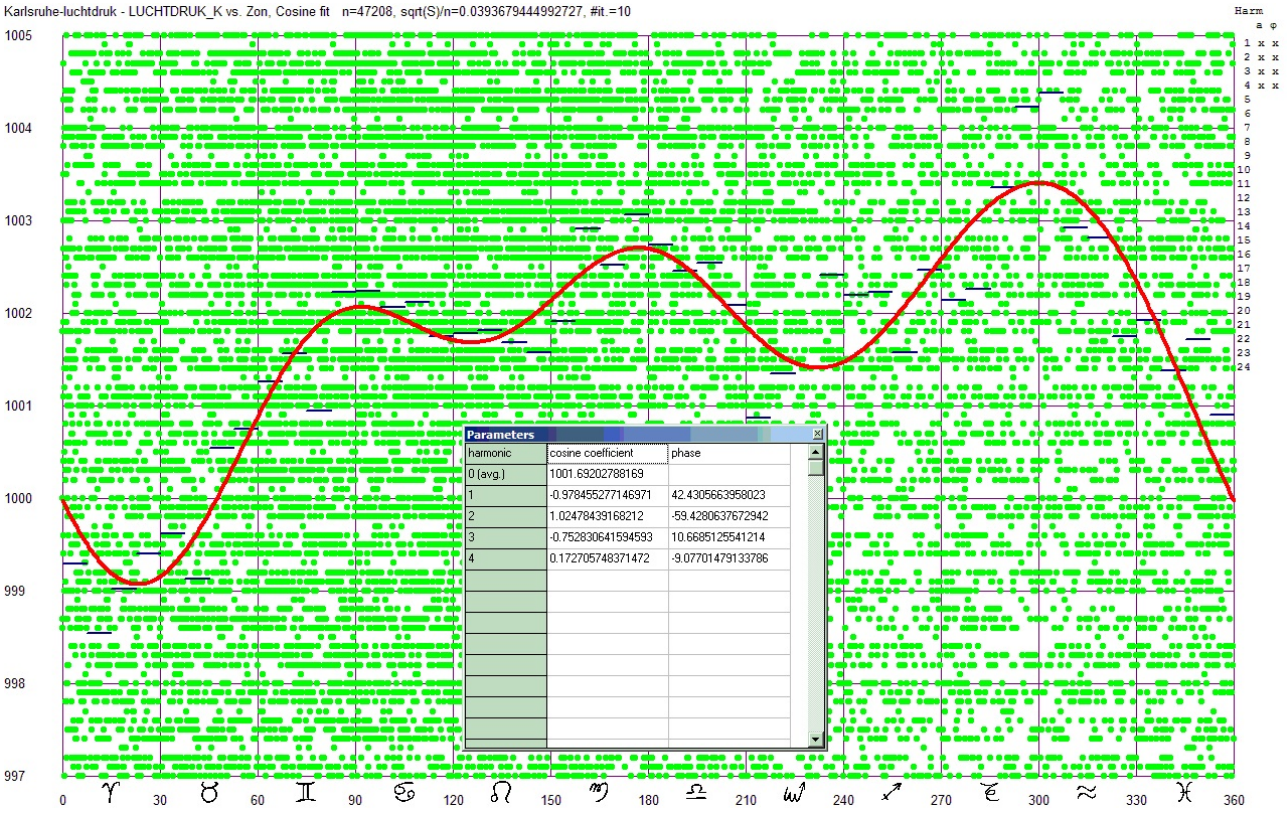
Stel dat er geen verband zou bestaan tussen de x-rij en de y-rij, dan zou de volgorde van de y's niet uitmaken, en zouden we met gelijk welke volgorde altijd een r^2 (of algemeen s) van dezelfde grootte-orde moeten bekomen, of beter: als we alle mogelijke verwisselingen van y-waarden doen en we bereken telkens de parameters en de s, dan zou de s van de echte meetgegevens niet mogen "opvallen" in het rijtje van de bekomen s'en.
Terug naar ons voorbeeld: als we 2 van de y-waarden omwisselen, bv.:

x_i	y_i
2005	109
2006	100
2007	120

dan krijgen we $r^2 = 0.30149...$, wat dus heel wat slechter is, en het is hier inderdaad zo op het zicht te zien dat het lineair verband niet meer duidelijk is. Alleen als we de eerste en de laatste y verwisselen, krijgen we een even goede korrelatie. Meestal (in 4 van de 6 gevallen) geeft verwisseling van de y-waarden een slechtere korrelatie, wat ons suggereert dat de echte meetgegevens "specialer" zijn dan de dooreengeschudde (er zit minder entropie in, zou je kunnen zeggen), als je met "speciaal" bedoelt "lineair verbonden".

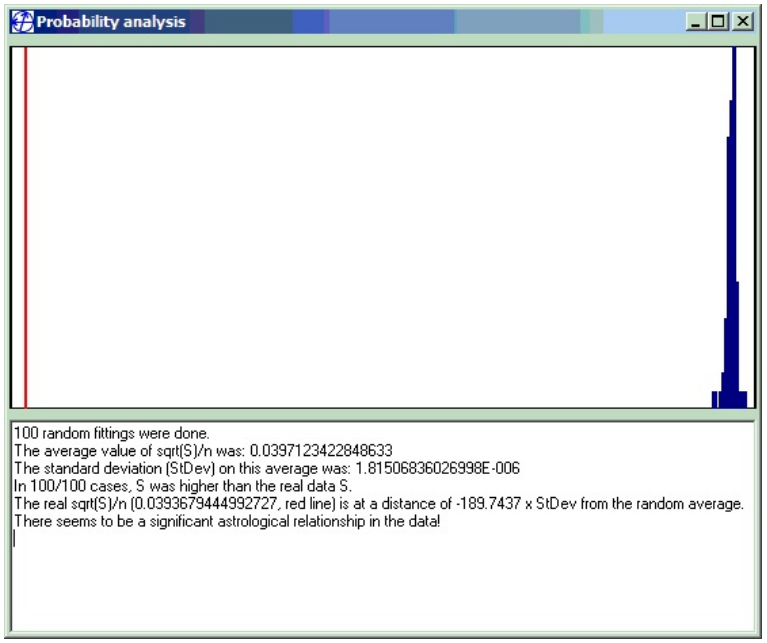
In de praktijk, met zeer veel gegevens, wordt het te tijdrovend om systematisch alle mogelijke verwisselingen uit te proberen. We kunnen dan bv. 100x alle y-waarden dooreenschudden, telkens s van de best passende f berekenen, en dan kijken hoeveel keren s groter was dan bij de echte gegevens, of kijken waar onze s ligt in de verdeling van de s'en en dan daarop een kansberekening loslaten.

Een praktisch voorbeeld: we beschikten over 47000 dagelijkse metingen van de luchtdruk in Karlsruhe. Vraag: "Is er een verband tussen de periode van het jaar en de luchtdruk?". Aangezien zulk een verband cyklisch zou moeten zijn, kunnen we het dus benaderen door (de eerste termen van) een Fourier-reeks. We namen de eerste 4 termen, d.w.z. een golf met periode van 1 jaar, 1/2 jaar, 1/3 jaar, en 1/4 jaar. Op de grafiek hieronder ziet u de metingen (bolletjes) en de best passende kurve van de gekozen vorm. (De positie in het jaar wordt hier gegeven als de ekliptische lengte van de zon, d.w.z. links = begin van de lente, rechts = einde van de winter.)



Het is onmogelijk om hier op het zicht te zeggen of dit verband nu “echt” is of niet. Er zit namelijk zeer veel “ruis” op de gegevens; er zijn zeker nog vele faktoren die de luchtdruk beïnvloeden.

Als we echter 100x alle luchtdrukwaarden dooreenschudden, verkrijgen we blijkbaar telkens een “best passende kurve” die minder goed past dan de echte. Op de volgende grafiek ziet u de verdeling van s bij de dooreengeschudde gegevens (blauw histogram) en een rode lijn die de positie van s bij de echte gegevens aangeeft.



De s van de echte gegevens ligt bijna 190 standaardafwijkingen links van het centrum van de s-verdeling, wat betekent dat de kans dat de kurve van de eerste grafiek toeval is, astronomisch klein is.
Mocht er geen verband bestaan, zou de rode lijn ergens in het histogram te zien zijn.

De grafieken werden gemaakt m.b.v. het programma Radix5, zie:
www.astrovdm.com/radix5nl.htm.
In het hulpbestand hiervan (gratis af te halen) staan nog meer voorbeelden.