

DE PRODUCT-MOMENT CORRELATIE-COEFFICIENT
EEN HISTORISCH OVERZICHT *

A. BOOMSMA

Introductie

Deze notitie geeft een kort overzicht van enkele feiten uit de geschiedenis van de product-moment correlatie-coëfficiënt. De lezer krijgt behalve de belangrijkste literatuurverwijzingen een globale indruk van de context waarbinnen deze coëfficiënt is ontstaan.

Volgens de samenstellers van een reeks artikelen over de geschiedenis van statistiek en waarschijnlijkheidstheorie is er op dit brede gebied erg weinig gepubliceerd (Pearson & Kendall, 1970). Gelukkig geldt dit niet voor de geschiedenis van de correlatie-coëfficiënt. Behalve een historisch artikel van K. Pearson (1920) is er een uitvoerige behandeling over het ontstaan van de correlatie-coëfficiënt te vinden bij K. Pearson (1930, hoofdstuk 14) en geeft Walker (1929) een algemeen historisch overzicht van meerdere correlatie-maten.

De rol van Bravais

In 1896 publiceerde Karl Pearson (1857-1936) een artikel dat nadien veel verwarring heeft gesticht, ondanks zijn pogingen later een en ander recht te zetten. Hij schreef toen (1896, p. 261): "The fundamental theorems of correlations were for the first time and almost exhaustively discussed by Bravais (*Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Mémoires par divers Savans, T. IX., Paris, 1846, pp. 255-332) nearly half a century ago. He deals completely with the correlation of two and three variables". In 1920 kent K. Pearson aan August Bravais een andere rol toe (p. 28): "Now I regret to say that nearly the whole of the above statements are hopelessly incorrect. Bravais has no claim, whatever, to supplant Francis Galton as the discoverer of the correlational calculus". In dit artikel stelt hij bijzonder duidelijk dat niemand de plaats van Francis Galton (1822-1911) kan innemen als de eerste die een 'volledig' begrip had van de fundamentele maat correlatie, zowel in praktisch als theoretisch opzicht.

* Eerder enigszins gewijzigd verschenen als Heymans-Bulletin, HB-203-EX, 1975, Groningen.

Oorsprong van de correlatie-coëfficiënt

Opgemerkt zij dat Galton een neef was van Charles Darwin (1809-1882), met wie hij zo nu en dan wetenschappelijk samenwerkte. Na zijn zestigste toen zijn belangrijkste wetenschappelijke bijdragen nog moesten worden geleverd, bestudeerde Galton bijna uitsluitend onderwerpen uit de erfelijkheids- en evolutieleer. De ontwikkeling van het begrip correlatie vond zijn oorsprong in praktische problemen die uit deze studie voortvloeiden.

Hij hield zich met name met de volgende vraagstukken bezig.

- a) Het probleem hoe de erfelijkheid gemeten moet worden. Galton probeerde een maat te vinden voor de gelijkheid tussen ouders en nakomelingen.
- b) Het probleem van de afhankelijkheid tussen twee variabelen, waarvan de waarden bij hetzelfde individu zijn vastgesteld. Hij was gefascineerd door het feit dat de grootte van twee verschillende organen gemeten bij hetzelfde levende wezen met elkaar varieerde; de organen waren ofwel beide groot ofwel beide klein, van onafhankelijkheid in grootte was geen sprake.

Wat Galton niet direct zag was dat de twee problemen zo nauw met elkaar verwant zijn, dat ze dezelfde oplossing in zich dragen. Eerst kwam hij evenwel op het begrip regressie, enkele jaren later op het begrip correlatie.

Een verdienstelijke samenvatting van Galton's statistische ontdekkingen wordt door Forrest (1974, hoofdstuk 14) gegeven.

Van "reversion" naar "regression"

Toen Galton in 1875 met assistentie van Darwin experimenteerde met het gewicht en de diameter van zaden van de lathyrus vond hij zijn eerste "regressie"-lijn.

Hij introduceerde zijn bevindingen tijdens een lezing aan de Royal Institution of Great Britain op 9 februari 1877. In deze lezing, getiteld "Typical laws of heredity in man", komt voor het eerst een numerieke maat r voor; deze letter r stond voor "reversion", hetgeen Galton later "regression" zou noemen. De oorsprong van het symbool voor de correlatie-coëfficiënt ligt dus bij de eerste letter van "reversion", niet van "regression". Wat Galton onder "reversion" verstond blijkt uit zijn eigen woorden (1877, p. 10): "Reversion is the tendency of the ideal mean filial type to depart from the parental type, reverting to what may be roughly and perhaps fairly described

as the average ancestral type. If family variability had been the only process in simple descent that affected the characteristics of a sample the dispersion of the race from its mean ideal type would indefinitely increase with the number of generations, but reversion checks this increase, and brings it to a standstill".

De volgende jaren verzamelde Galton materiaal voor zijn verdere onderzoek naar regressie en erfelijkheid. Een belangrijke publicatie verscheen in 1885, waarin hij zijn werk over regressie in uitvoerige vorm uiteenzet. In zijn artikel vindt men een correlatietabel (een tabel die de gegevens bevat waaruit de correlatie-coëfficiënt wordt berekend) en een diagram van de bijbehorende twee regressielijnen.

We zullen niet aangeven hoe Galton zijn ideeën over regressie uitwerkte en verder ontwikkelde. Het is voldoende te constateren dat zijn ontdekking van het begrip regressie voorafging aan dat van de correlatie-coëfficiënt.

Van "co-relation" naar "correlation"

Het duurde tot 1888 voordat Galton zich realiseerde dat de twee hiervoor omschreven problemen identiek zijn, als tenminste iedere variabele wordt gemeten in de eenheid van zijn eigen standaardafwijking. Onder deze conditie is de helling van de regressielijn gelijk aan de product-moment correlatie-coëfficiënt.

Deze ontdekking werd openbaar gemaakt op 20 december 1888 in de voordracht van zijn artikel "Co-relations and their measurement chiefly from anthropometric data" voor de Royal Society of London. Galton kwam toen met de volgende omschrijving (p. 135): "Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction". Op dezelfde pagina blijkt zijn begrip van de nieuwe associatiemaat. "It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to common causes, the co-relation would be nil. Between these two extremes are an endless number of intermediate cases, and it will be shown how the closeness of co-relation in any particular case admits of being expressed

by a single number".

Het artikel bevat een aantal correlatie-coëfficiënten gebaseerd op metingen verricht bij 350 volwassen mannen. Als voorbeelden van de eerste correlaties door Galton berekend, noemen we: r (lengte onderarm, lichaamslengte) = 0.80, r (kniehoogte, lichaamslengte) = 0.90, r (hoofdbreedte, hoofdlangte) = 0.45.

We geven geen gedetailleerde beschrijving van de manier waarop Galton correlaties berekende; de lezer wordt daarvoor verwezen naar publicaties van Galton (1889), K. Pearson (1920, 1930) en Forrest (1974). Een aantal opmerkingen over zijn correlatie-coëfficiënt zijn evenwel van belang.

Allereerst berekende hij de correlaties niet op de manier zoals dat tegenwoordig veelal gebeurt, waarbij de correlatie-coëfficiënt r gelijk is aan het gemiddelde product van de gestandaardiseerde scores (zogenaaamde z -scores). Galton gebruikte in zijn formules geen gemiddelden maar medianen, terwijl hij werkte met interkwartiel afstanden in plaats van met standaardafwijkingen.

Ten tweede is het duidelijk dat hij destijds nog niet het bestaan en het belang van negatieve correlaties onderkende. Het was Weldon (1892) die negatieve correlaties introduceerde.

Een derde kenmerk van zijn behandeling van correlaties is dat hij steeds uitgaat van een bivariaat normale verdeling van de variabelen. Karl Pearson (1930, p. 5-6) schreef daarover het volgende: "He started like Quetelet from the normal curve as describing the deviations of a population or of any selected population, e.g. that of an array of offspring from the parent of given character. He did not start with a general definition of correlation and see whither that would lead him. His justification was that he was dealing with anthropometric characters or measurements on living forms whose deviations from type approximately followed this special law of distribution. Thus he naturally selected a straight regression line, and the constant variability for all arrays of one character for a given value of a second. It was perhaps, best for the progress of correlational calculus that this simple special case should be promulgated first; it is so easily grasped by the beginner. But it has had the disadvantage that certain branches of science, as psychology for example, have rarely got further, and, without taking the trouble to apply tests, adopt linear regression and homoscedasticity where it is quite inappropriate. It is interesting to note how the history of the spread of knowledge follows with halting steps the history of its discovery".

Veel van het werk van Karl Pearson gaat over correlaties en afhankelijkheden tussen variabelen die niet gebaseerd zijn op de bivariate normaliteitsassumptie. (Voor een overzicht van zijn werk verwijzen we naar zijn zoon Egon S. Pearson (1938) en Walker (1958))

Hoewel ook hij in hoofdzaak werkte met antropometrische gegevens was hij praktisch genoeg om het bestaan van "niet-normale" gegevens te onderkennen. Zo is zijn hele systeem van frequentie curven (Pearson, 1895) ontstaan uit de overweging dat niet alles normaal, laat staan bivariaat normaal verdeeld is. Iets meer dan een halve eeuw terug schreef hij (1920, p. 44): "This seems to me the desideratum of the theory of correlations at the present time: the discovery of an appropriate system of surfaces, which will give bi-variate skew frequency. We want to free ourselves from the limitations of the normal surface, as we have from the normal curve of errors". Helaas is er sedertdien in dit opzicht weinig vooruitgang geboekt. De vooruitgang die sinds Galton wel werd geboekt wordt helaas niet overal toegepast; zo wordt in veel gevallen bij het gebruik van correlatie-coëfficiënten zonder meer bivariate normaliteit verondersteld.

K. Pearson

De grote invloed van Galton's ideeën over correlatie kwam vooral na publicatie van zijn boek *Natural Inheritance*, in 1889. In hoofdstuk 5 van dit boek geeft Galton een volledige behandeling van zijn werk over regressie en correlatie dat besloten ligt in zijn eerdere geschriften van 1877 tot 1888. De invloed van *Natural Inheritance* moge blijken uit een reactie van Karl Pearson: "...there was a category broader than causation; namely correlation, of which causation was only the limit, and that this new conception of correlation brought psychology, anthropology, medicine and sociology in large parts into the field of mathematical treatment. It was Galton who first freed me from the prejudice that sound mathematics could only be applied to natural phenomena under the category of causation". (geciteerd bij E.S. Pearson, 1938, p. 19)

Volgens K. Pearson (1930, p. 57) ligt bij de publicatie van *Natural Inheritance* het ontstaan van de school van Galton. Het bestuderen van de correlatie-coëfficiënt door K. Pearson, W.F.R. Weldon en F.Y. Edgeworth was er het gevolg van. (Voor de bijdragen van Weldon en Edgeworth verwijzen we naar K. Pearson (1920) en Walker (1929))

De methode van correlatieberekening door Galton was vrij grof en voor

verbetering vatbaar. Bovendien had K. Pearson op verschillende punten kritiek op Natural Inheritance; naar zijn mening interpreteerde Galton bij erfelijkheidsproblemen regressie verkeerd, omdat hij zich beperkte tot het begrip regressie met twee variabelen en geen kennis had van multiple regressie en multiple correlatie.

Zo bleven er tal van problemen liggen op het gebied van correlatie en regressie. Het is vooral K. Pearson geweest die rond de eeuwwisseling een grote bijdrage heeft geleverd tot de oplossing van die problemen.

Hoe sterk Pearson daarbij werkte in de traditie van Galton, wordt bevestigd door E.S. Pearson (1938, p. 26-27). "The main purpose of all this work was the development and application of statistical methods for the study of heredity and evolution; it would certainly be wrong to think of the Pearson of this period as concerned with the development of statistical theory for its own sake. We may trace two important objectives which he had in view: the testing of the adequacy of Galton's Law of Ancestral Heredity and the development of theory that would throw light on the influence of selection on the correlation and variability of organs".

In 1896 publiceert K. Pearson een artikel waarin de correlatie-coëfficiënt in zijn huidige vorm verschijnt. Op bladzijde 264-265 geeft hij het bewijs dat de product-moment uitdrukking voor r

$$\Sigma (xy) / (\Sigma x^2 \Sigma y^2)^{1/2} \quad (1)$$

waarin x en y afwijkingen zijn van hun steekproefgemiddelde, de "beste" schatter voor de populatie correlatie-coëfficiënt ρ is.

Naast de grote algemene betekenis die Pearson had voor de ontwikkeling van de correlatie-coëfficiënt ligt hier de reden dat tegenwoordig het gebruik van de term Pearson product-moment correlatie in zwang is. Als het enige zin zou hebben statistische grootheden namen van hun ontdekkers mee te geven, zou het om historische redenen juist zijn r de Galton-Pearson correlatie-coëfficiënt te noemen.

Student

Met de definitie van de correlatie-coëfficiënt was niet direct de steekproevenverdeling van r bekend. Het was W.S. Gosset die onder zijn bescheiden pseudoniem voor het geval $\rho = 0$ een benadering van de steekproevenverdeling van r voorstelde bij normaal verdeelde variabelen (1908-1909).

Soper (1913) leidt benaderingen af voor de standaardafwijking en de gemiddelde waarde van r bij steekproeven ter grootte n uit een bivariaat normale verdeling met correlatie ρ .

R.A. Fisher

Twee belangrijke artikelen van R.A. Fisher sluiten de hoofdlijnen in de geschiedenis van de correlatie-coëfficiënt af. In 1915 komt Fisher met de exacte steekproevenverdeling van r , terwijl hij in 1921 het belang van de zogenaamde Fisher z -transformatie, $z = \tanh^{-1} r = \frac{1}{2} \log \frac{1+r}{1-r}$, aantoonst.

Hoewel er na 1921 nog wel enkele belangwekkende studies over de correlatie-coëfficiënt zijn verschenen, waarbij bijvoorbeeld gedacht kan worden aan een bijdrage van H. Hotelling (1953), was het meest belangrijke werk met de afleiding van de steekproevenverdeling van r en de presentatie van de z -transformatie voltooid.

Literatuur

- Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in small samples from an indefinitely large population. Biometrika, 10, 507-521.
- Fisher, F.A. (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1 (4), 3-32.
- Forrest, D.W. (1974). Francis Galton. The life and work of a Victorian genius. London: Elek.
- Galton, F. (1877). Typical laws of heredity in man. Address to the Royal Institution of Great Britain. (Typical laws of heredity. Proceedings of the Royal Institution, 8, 282-301)
- Galton, F. (1885). Regression towards mediocrity in hereditary stature. J. Anthropological Institute of Great Britain and Ireland, 15, 246-263.
- Galton, F. (1888). Co-relations and their measurement chiefly from anthropometric data. Proceed. Roy. Soc. London, 45, 135-145.
- Galton, F. (1889). Natural Inheritance. London: MacMillan.

- Hotelling, H. (1953) New light on the correlation coefficient and its transforms. J. Roy. Statist. Soc., B, 15, 193-232.
- Pearson, E.S. (1938). Karl Pearson. An appreciation of some aspects of his life and work. Cambridge University Press.
- Pearson, E.S. & Kendall, M.G. (Eds.) (1970). Studies in the history of statistics and probability. London: Griffin.
- Pearson, K. (1895). Mathematical contributions to the theory of evolution. II. Skew variation in homogeneous material. Phil. Trans. Roy. Soc., 186, A, 343-414.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Phil. Trans. Roy. Soc., 187, A, 253-318.
- Pearson, K. (1920). Notes on the history of correlation. Biometrika, 13, 25-45.
- Pearson, K. (1914, 1924, 1930). The life, letters and labours of Francis Galton. Vol. I, II, III. Cambridge University Press.
- Soper, H.E. (1913) On the probable error of the correlation coefficient to a second approximation. Biometrika, 9, 91-115.
- Student (1908-1909) Probable error of a correlation coefficient. Biometrika, 6, 302-310.
- Walker, H.M. (1929). Studies in the history of statistical method. Baltimore: Williams & Wilkins.
- Walker, H.M. (1958). The contributions of Karl Pearson. J. Amer. Statist. Assoc., 53, 11-22.
- Weldon, W.F.R. (1892). Certain correlated variations in Crangon vulgaris. Proceed. Roy. Soc. London, 51, 2-21.