A Correlation for the 21st Century

Terry Speed

Science **334**, 1502 (2011);
DOI: 10.1126/science.1215894

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

CORRECTED 13 JANUARY 2012; SEE LAST PAGE

(*1*). As a result, HCQ cancer trials have been launched. Autophagy inhibition with HCQ has been observed in patient samples, and some preliminary results are encouraging. For instance, a phase I trial of the drug temsirolimus with HCQ showed stabilized tumor growth in 73% of patients with metastatic melanoma (*15*), whereas temsirolimus alone produced a 0% stable disease rate (*16*). The findings by Michaud *et al.* suggest that trials combining cancer therapies with autophagy inhibitors should incorporate measurements of host immunity before and during treatment with autophagy inhibitors in mouse models and patients. Integrating clinic and lab findings will be crucial to understanding the complex role of autophagy in tumor immunity.

References and Notes

1. R. K. Amaravadi *et al.*, *Clin. Cancer Res.* **17**, 654 (2011).
2. M. Michaud *et al.*, *Science* **334**, 1573 (2011).
3. F. Ghiringhelli *et al.*, *Nat. Med.* **15**, 1170 (2009).
4. M. Z. Noman *et al.*, *Cancer Res.* **71**, 5976 (2011).
5. K. P. Olive *et al.*, *Science* **324**, 1457 (2009).
6. G. L. Beatty *et al.*, *Science* **331**, 1612 (2011).
7. N. J. Ives *et al.*, *J. Clin. Oncol.* **25**, 5426 (2007).
8. S. Yang *et al.*, *Genes Dev.* **25**, 717 (2011).
9. X. H. Ma *et al.*, *Clin. Cancer Res.* **17**, 3478 (2011).
10. R. Lazova *et al.*, *Clin. Cancer Res.* 10.1158/1078-0432. CCR-11-1282 (2011).
11. X. Qu *et al.*, *J. Clin. Invest.* **112**, 1809 (2003).
12. H. Wei *et al.*, *Genes Dev.* **25**, 1510 (2011).
13. J. Liu *et al.*, *Cell* **147**, 223 (2011).
14. J. Y. Guo *et al.*, *Genes Dev.* **25**, 460 (2011).
15. S. L. Algazy *et al.*, in *Proceeding of the 102nd Annual Meeting of the AACR*, Orlando, FL, 2 to 6 April 2011 (AACR, Philadelphia, PA, 2011), abstract 4500.
16. K. Margolin *et al.*, *Cancer* **104**, 1045 (2005).
17. R.K.A. is an investigator on HCQ trials.

10.1126/science.1216428

MATHEMATICS

A Correlation for the 21st Century

A novel statistical approach has been developed that can uncover nonlinear associations in large data sets.

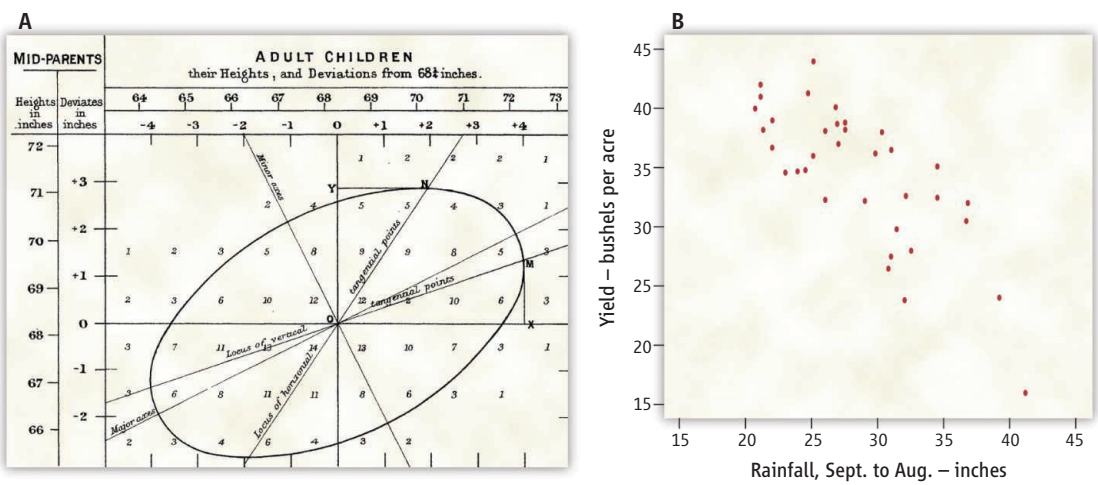
Terry Speed

Most scientists will be familiar with the use of Pearson's correlation coefficient r to measure the strength of association between a pair of variables: for example, between the height of a child and the average height of their parents ($r \approx 0.5$; see the figure, panel A), or between wheat yield and annual rainfall ($r \approx 0.75$, panel B). However, Pearson's r captures only linear association, and its usefulness is greatly reduced when associations are nonlinear. What has long been needed is a measure that quantifies associations between variables generally, one that reduces to Pearson's in the linear case, but that behaves as we'd like in the nonlinear case. On page 1518 of this issue, Reshef *et al.* (*1*) introduce the maximal information coefficient, or MIC, that can be used to determine nonlinear correlations in data sets equitably.

The common correlation coefficient r was invented in 1888 by Charles Darwin's half-cousin Francis Galton (*2*). Galton's method for estimating r was very different from the one we use now, but was amenable to hand calculation for samples of up to 1000 individuals. Francis Ysidro Edgeworth and later Karl Pearson gave us the modern formula for estimating r, and it very definitely required a manual or electromechanical calculator to convert 1000 pairs of values into a correlation coefficient. In marked contrast, the MIC requires a modern digital computer for its calculation; there is no simple formula, and no-one could compute it on any calculator. This is another instance of computer-intensive methods in statistics (*3*).

It is impossible to discuss measures of association without referring to the concept of independence. Events or measurements are termed probabilistically independent if information about some does not change the probabilities of the others. The outcomes of successive tosses of a coin are independent events: Knowledge of the outcomes of some tosses does not affect the probabilities for the outcomes of other tosses. By convention, any measure of association between two variables must be zero if the variables are independent. Such measures are also called measures of dependence. There are several other natural requirements of a good measure of dependence, including symmetry (*4*), and statisticians have struggled with the challenge of defining suitable measures since Galton introduced the correlation coefficient. Many novel measures of association have been invented, including rank correlation (*5*, *6*); maximal linear correlation after transforming both variables (*7*), which has been rediscovered many times since; the curve-based methods reviewed in (*8*); and, most recently, distance correlation (*9*).

To understand where the MIC comes from, we need to go back to Claude Shan-



CREDIT: (A) F. GALTON

Finding correlations. (**A**) Association between the height of adult children and the average height of their parents [adapted from (*11*)]. (**B**) Association between wheat yield and annual rainfall [adapted from (*12*)].

Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville VIC 3052, Australia, and Department of Statistics, University of California, Berkeley, CA 94720, USA. E-mail: terry@stat.berkeley.edu

Published by AAAS

non, the founder of information theory. Shannon defined the entropy of a single random variable, and laid the groundwork for what we now call the mutual information, MI, of a pair of random variables. This quantity turns out to be a new measure of dependence and was first proposed as such in 1957 (*10*). Reshef *et al*.'s MIC is the culmination of more than 50 years of development of MI.

What took so long, and wherein lies the novelty of MIC? There were three difficulties holding back MI's acceptance as the right generalization of the correlation coefficient. One was computational. It turns out to be surprisingly tricky to estimate MI well from modest amounts of data, mainly because of the need to carry out two-dimensional smoothing and to calculate logarithms of proportions. Second, unlike the correlation coefficient, MI does not automatically come with a standard numerical range or a ready interpretation of its values. A value of $r = 0.5$ tells us something about the nature of a cloud of points, but a value of MI = 2.2 does not. The formula $[1 - \exp(-2\mathrm{MI})]^{1/2}$ in (*10*) satisfies all the requirements for a good measure of dependence, apart from ease of computation, and ranges from 0 to 1 as we go from independence to total dependence. But Reshef *et al*. wanted more, and this takes us to the heart of MIC. Although r was introduced to quantify the association between two variables evident in a scatter plot, it later came to play an important secondary role as a measure of how tightly or loosely the data are spread around the regression line(s). More generally, the coefficient of determination of a set of data relative to an estimated curve is the square of the correlation between the data points and their corresponding fitted values read from the curve. In this context, Reshef *et al*. want their measure of association to satisfy the criterion of equitability, that is, to assign similar values to "equally noisy relationships of different types." MI alone will not satisfy this requirement, but the three-step algorithm leading to MIC does.

Is this the end of the Galton-Pearson correlation coefficient r? Not quite. A very important extension of the linear correlation r_{XY} between a pair of variables X and Y is the partial (linear) correlation $r_{XY.Z}$ between X and Y while a third variable, Z, is held at some value. In the linear world, the magnitude of $r_{XY.Z}$ does not depend on the value at which Z is held; in the nonlinear world, it may, and that could be very interesting. Thus, we need extensions of MIC(X,Y) to MIC(X,Y|Z). We will want to know how much data are needed to get stable estimates of MIC, how susceptible it is to outliers, what three- or higher-dimensional relationships it will miss, and more. MIC is a great step forward, but there are many more steps to take.

References

1. D. N. Reshef *et al*., *Science* **334**, 1518 (2011).
2. S. Stigler, *Stat. Sci.* **4**, 73 (1989).
3. P. Diaconis, B. Efron, *Sci. Am.* **248**, 116 (1983).
4. A. Rényi, *Acta Math. Hung.* **10**, 441 (1959).
5. C. Spearman, *Am. J. Psychol.* **15**, 72 (1904).
6. M. G. Kendall, *Biometrika* **30**, 81 (1938).
7. H. O. Hirschfeld, J. Wishart, *Math. Proc. Camb. Philos. Soc.* **31**, 520 (1935).
8. P. Delicado, M. Smrekar, *Stat. Comput.* **19**, 255 (2009).
9. G. J. Székely, M. L. Rizzo, *Ann. Appl. Stat.* **3**, 1236 (2009).
10. E. H. Linfoot, *Inf. Control* **1**, 85 (1957).
11. F. Galton, *J. Anthropol. Inst. Great Brit. Ire.* **15**, 246 (1886).
12. R. A. Fisher, in *Statistical Methods for Research Workers* (Oliver & Boyd, Edinburgh, 1925), p. 34.

10.1126/science.1215894

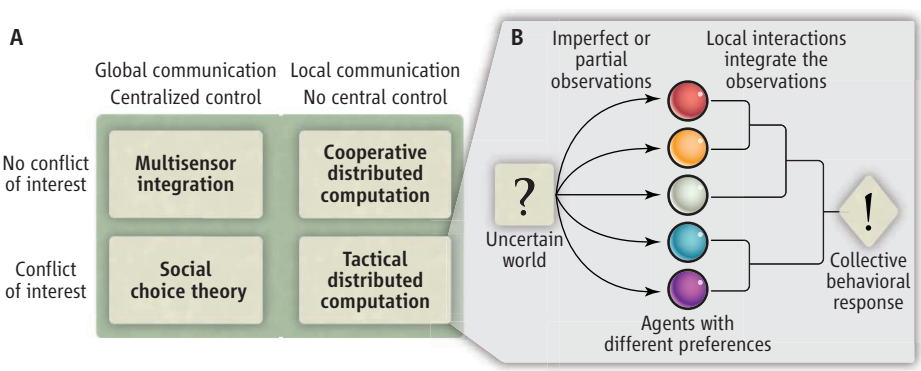
PSYCHOLOGY

Can Ignorance Promote Democracy?

When a group needs to reach a consensus decision, uninformed members can help to reduce the influence of a manipulative minority.

Jevin D. West[1] and Carl T. Bergstrom[1,2]

Ideas are like fire, observed Thomas Jefferson in 1813—information can be passed on without relinquishing it (*1*). Indeed, the ease and benefit of sharing information select for individuals to aggregate into groups, driving the buildup of complexity in the biological world (*2*, *3*). Once the members of some collective—whether cells of a fruit fly or citizens of a democratic society—have accumulated information, they must integrate that information and make decisions based upon it. When these members share a common interest, as do the stomata on the surface of a plant leaf (*4*), integrating distributed information may be a computational challenge. But when individuals do not have entirely coincident interests, strategic problems arise. Members of animal herds, for example, face a tension between aggregating information for the benefit of the herd as a whole, and avoiding manipulation by self-interested individuals in the herd. Which collective decision procedures are robust to manipulation by selfish players (*5*)? On page 1578 of this issue, Couzin *et al*. (*6*) show how the presence of uninformed agents can promote democratic outcomes in collective decision problems.



Distributed information processing. (**A**) Research in this domain comprises four areas: multisensor integration (*12*), social choice theory (*13*), cooperative distributed computation (*14*), and tactical distributed computation (*5*). The Couzin *et al*. study lies in the lower right quadrant, where the challenges of both social choice and distributed computation must be solved. (**B**) In this schematic of tactical distributed computation, an uncertain world is observed imperfectly by agents with different preferences. By means of local interactions, they aggregate the information and preferences to arrive at a collective decision.

[1]Department of Biology, University of Washington, Box 351800, Seattle, WA 98195–1800, USA. [2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA. E-mail: jevinw@u.washington.edu, cbergst@u.washington.edu

ERRATUM

Post date 13 January 2012

Perspectives: "A correlation for the 21st century" by T. Speed (16 December 2011, p. 1502). Owing to a production error, the name J. Wishart was mistakenly included as a coauthor in reference 7. H. O. Hirschfeld is the sole author.