Are the Data Worth Owning?

Author(s): J. Ross Macdonald
Source: *Science*, New Series, Vol. 176, No. 4042 (Jun. 30, 1972), p. 1377
Published by: American Association for the Advancement of Science
Stable URL: http://www.jstor.org/stable/1734564
Accessed: 01/05/2009 11:05

30 June 1972, Volume 176, Number 4042

SCIENCE

Are the Data Worth Owning?

Amitai Etzioni has recently (14 April 1972, page 121) raised the question of who should ultimately own the data. He points out that, since data is (sic) often lost or becomes inaccessible, agencies that finance data collection or preparation should require that it be made available to others by deposition in a data bank or library.

A more fundamental question than who should own the data is, Are the data even worth owning? Unfortunately, the answer is usually an embarrassing and costly "No" across the entire spectrum of research. The problem usually lies in lack of knowledge about the trustworthiness of the data. Measures of uncertainty are usually not given at all; even when they are, they are themselves generally untrustworthy. Lancelot Hogben has stated that "less than one percent of research workers clearly apprehend the rationale of statistical techniques they commonly invoke."

Further, David Lide, head of the National Standard Reference Data System of the National Bureau of Standards, estimates that from 50 to over 90 percent of the published raw data available for producing trustworthy, evaluated results for the physical properties of scientific materials cannot, in fact, be used for this important purpose. A good illustration of the basic difficulty has been given by the late W. J. Youden of NBS. He states that, of 15 observations of the mean distance to the sun published from 1895 to 1961, each worker's estimated value is outside the uncertainty limits set by his immediate predecessor.

Both systematic and random errors occur in all experimental situations. They should be estimated, discussed, and cited separately, as Churchill Eisenhart has pointed out. Ideally, systematic errors should be estimated by independently measuring the quantity in question with a different apparatus, preferably one that operates on a different principle from that of the original apparatus. One should strive to make the estimated maximum systematic error comparable to or smaller than the estimated root-mean-square random error of the experiment. When it is impractical to obtain independent estimates of the systematic error, a good rule of thumb is to multiply one's best estimate of it by a factor of 3.

Even when estimates of individual errors (deviations) are calculated, it is unusual for the experimenter to check the deviations for stochastic independence and to state the result of such a check. Rarely indeed does one find mention of the statistical distribution that the deviations appear to follow. Without such knowledge, however, one cannot assess the meaningfulness of such important derived quantities as ordinary confidence limits and standard deviations. When individual measurements are to be analyzed by such techniques as least squares, one seldom finds experimenters replicating the individual measurements closely enough to obtain trustworthy estimates of uncertainties for use in weighted least squares. Yet only thus can one verify the assumptions implicit in even unweighted least squares. Nonlinear least squares analysis is becoming much more common these days, but one never finds it shown that the bias in parameter estimates introduced by this estimation technique is safely smaller than the sampling error. Finally, there are almost always random errors present in the values of all variables measured (except in whole number cases), not just in the "dependent" variable, as is usually assumed in ordinary least squares analysis. Although a generalized least squares technique is necessary and available to handle such situations, it is hardly ever used—nor is the need for it usually recognized.

Clearly, much further education in data analysis, presentation of results, and the need to call in a statistician is necessary before a high proportion of published data can be properly used for more than qualitative purposes.—J. ROSS MACDONALD, *Chairman, Numerical Data Advisory Board, National Research Council.*