MARIANNE COOPER and HERBERT W. COOPER

Interpreting Results of Statistical Studies

Statistical analysis is a tool to be used by librarians only after critical decisions have been made about the nature of their libraries' operations. It is shown that different conclusions might be drawn from identical data by presupposing different, but equally plausible, relationships between the variables. Different conclusions might also result from analyzing subsets of data rather than the sample as a whole. The role of the librarian as a basic decision-maker in statistical studies is emphasized by presenting alternative interpretations of a hypothetical set of data.

LIBRARIANS, LIKE OTHER professional scholars, are making increased use of statistical techniques to analyze data. Their purpose is to make comparisons or forecasts that are as free as possible from subjective factors or preconceived notions. In order to use statistical techniques effectively, however, librarians must consider two points very carefully. Both are related to the relative positions of the librarian and the tool called statistical analysis.

The first point is that the librarian must decide what relationships can reasonably be expected to exist between the variables of interest. *After* the librarian makes this decision, statistical analysis can be used to evaluate the most likely values of various constants.

This sequence of events is nicely illustrated in a recent paper by Reichard and Orsagh.[1] They collected values of five variables for each of approximately three hundred institutions. The data were: expenditures for current acquisitions, holdings, number of undergraduates, number of graduate students, and number of faculty members. The investigators then decided, *a priori,* that they would use equations of the following form to correlate the data.[2]

$$E = C_1 + C_2U + C_3G + C_4F \quad (1)$$

After this decision was made, statistical analysis was used to provide the best values of the constants C_1, C_2, C_3, and C_4. The point to be emphasized is that the investigator provides the basic form of the equation, and statistical analysis provides the values of the constants.

Equation 1 is the mathematical equivalent of the following statements. "The increases in expenditures are directly proportional to the increases in the number of undergraduates, number of graduate students, and number of faculty members. It costs C_2 dollars to add one undergraduate, C_3 dollars to add one graduate student, and C_4 dollars to add one faculty member, in all schools, at

[1] Edwin W. Reichard and Thomas J. Orsagh, "Holdings and Expenditures of U.S. Academic Libraries," *College and Research Libraries,* XXVII (November 1966), 479-87.

[2] E, U, G, F refer respectively to expenditures for current acquisitions, and number of undergraduates, graduates, and faculty.

Dr. Herbert Cooper is Director of Research for Alcorn Combustion Co. in New York. Mrs. Cooper is a DLS student at Columbia University.

all enrollment and faculty levels." While this is certainly a plausible assumption, it leads to the conclusion that the undergraduates have, if anything, a somewhat negative influence on expenditures and that it costs seven times as much to add a faculty member as it does to add a graduate student.

Other conclusions, however, might be drawn from the same data if a slightly different viewpoint were initially taken. In a library there must, of course, be a bare minimum of equipment and personnel present to serve even one patron. Once these minima are present, two, three, or more can be served with little increase in expenditure. This situation seems to be analogous to that of a chemical plant where a certain minimum amount of equipment and manpower is required to produce even a trickle of product. Again after these minimum facilities are present, production can be increased, up to a point, by relatively slight additional expenditures. In both the library and the chemical plant the additional expenses are less than proportional to the additional "throughput." To pursue this analogy we note that costs of chemical plants (and also many other items) are related to their size by an equation of the form:

$$\text{cost} = K\ (\text{size})^n \qquad (2)$$

where n is typically between 0.4 and 0.8, depending on the type of plant and, somewhat, on its size range. Equation 2 implies that the *per-cent* increase in cost is related to the *per-cent* increase in size. If the exponent n is less than 1.0 the per-cent cost increase is less than proportional to the per-cent size increases. We would therefore expect, in view of the above, that equations of the following form might be more realistic correlators.

$$E = K\ U^A\ G^B\ F^C \qquad (3)$$

Since this equation differs fundamentally from the linear ones presented in the aforementioned article it is possible that different conclusions might result. Whether this would, in fact, be the case could be established by making simple logarithmic transformations and re-analyzing the data.

This approach was tried with the hypothetical numbers listed in Table 1 of this article, constructed to give results similar to those of Reichard and Orsagh. Multivariate analysis of the type they used led to the following equation.

$$E = 5321 - 8.7U + 116G + 188F \qquad (4)$$

The coefficients indicate that it would cost \$116 to add a graduate student, \$188 to add a faculty member, but –\$8.70 (*i.e.*, a credit) to add an undergraduate *if equation 2 is accepted.* If, however, the data are analyzed by presupposing the percentage type of relationship implied by Equation 3, one obtains the following.

$$E = 202\ U^{0.157}\ G^{0.334}\ F^{0.5} \qquad (5)$$

This indicates that expenditures must be increased by 1.5 per cent, 3.2 per cent, and 4.8 per cent for a 10 per cent increase in the number of undergraduates, graduate students, and faculty, respectively, *if equation 3 is accepted.*

Which equation should be accepted? Unless the "coefficient of multiple determine" is much greater for one of the equations the librarian must use his experience, judgment, and knowledge of library procedures, to make this decision. This question always arises in empirical studies where causes and effects are not investigated. The first point, then, is that it is very important to examine the nature of the equations chosen to represent the data.

The second point for librarians to consider is that every library situation is, to some extent, special and unique. Reichard and Orsagh point out that one must use quantifiable variables, while recognizing that many other factors will exert

TABLE 1
HYPOTHETICAL USER AND EXPENDITURE DATA

Library	U	G	F	E
1	500	100	50	15,000
2	1,000	180	90	38,000
3	1,200	180	150	43,000
4	1,400	200	120	43,000
5	1,400	220	190	55,000
6	4,000	800	350	125,000
7	4,400	700	500	145,000
8	4,500	750	450	150,000
9	5,500	900	600	165,000
10	6,000	1,500	800	280,000

their influences. For example, small college libraries are likely to have different organizational patterns from those of large university libraries; while their budgets will differ, their service to clientele might be equally effective.

Consider a hypothetical university serving thirteen hundred undergraduates, two hundred graduate students, and one hundred and thirty faculty members. Its library expenditures would be estimated at $41,600 (from Equation 4) or $41,500 (from Equation 5), if it were considered to be an average library belonging to a population of which the ten (hypothetical) samples listed in Table 1 are representative. Examination of the data, however, suggests two populations; small libraries (numbers 1 to 5) and large libraries (numbers 6 to 10). If a multivariate analysis of the first five samples, only, is performed one obtains the following.

$$E = -13294 - 6.62U + 270G + 97F \quad (6)$$

and

$$E = 20\ U^{-0.16}\ G^{1.431}\ F^{0.265} \quad (7)$$

These equations lead to predictions of $44,700 and $45,800, respectively, for the library, considered as a "small" library. It is noted, parenthetically, that the relative influence of undergraduates, graduate students, and faculty is different in "small" institutions. It is emphasized, however, that Table 1 is hypothetical and has been prepared only to illustrate the points herein discussed. The essence of the second point is that by stratifying data one might obtain a different but more useful comparison. The librarian must furnish the basis of comparison. This is a matter requiring the librarian's expertise, experience, and insights.

Both of the points discussed above can be summarized as follows.

1. The forms of types of relationships assumed to relate the variables are chosen before statistical analysis is undertaken. These relationships must be examined critically in all cases.
2. Librarians must decide which data are to be analyzed, which form useful or natural subgroups, etc. These are questions relating to librarianship and are not statistical questions.

■■