AN INFORMATION THEORY APPLICATION TO MYCENAEAN POTTERY

Over 35 years ago the distinguished Swedish archaeologist Arne Furumark published his major work on Mycenaean pottery. In his study Furumark analyzed over 3,000 ceramic artifacts of Mycenaean origin and elaborated a typology of 103 forms, 336 types, and a typology of 80 decorative motifs with which many of these pots were painted. Although considerably out of date and in need of major revision, Furumark's work has remained the point of departure for most subsequent analyses of Mycenaean pottery.

During the summer and autumn of 1976, a pilot project was undertaken in the Institute for Advanced Study at Princeton to explore the feasibility of rendering Furumark's classification scheme amenable to computer treatment. About 3,400 artifacts cited both in F's Catalogue of Vessels and elsewhere in his study were selected and data concerning each artifact were coded and punched on 80-column punchcards. This information was organised in the following fashion:

(i) Identification of the evidence - material of fabrication, place found, cultural assemblage to which artifact belongs, relative chronological dating of the object (if ascertainable);

(ii) Weighting of the evidence - doubts about provenience, type, or archaeological context;

(iii) Classification of the evidence - Furumark's typologies of form and of decorative motifs;

(iv) Sources of data about the evidence - bibliographical references (if any); if unpublished, museum or collection accession number of find (if known);

(v) Type serial number - Furumark's type example numbering system was employed as a control reference number.

The feasibility study showed that, of the principal data categories used by Furumark, only his syntactical classification scheme did not lend itself too readily to digitized format. The initial results of this feasibility study could comprise a point of departure for constructing a completely revised, new typology.

A computerized system like the one described above provides the necessary data base for statistical analyses. The following account illustrates some of the ways in which this data base may be analytically explored.

Initially, the distribution of pottery was studied as conditioned by five variables: Type, Site, Context, Decorative Motif, and Date. These distributions were elaborated in tables, commonly known as multidimensional contingency tables (Ku, Varner, Kullback 1971), which can be readily generated in a convenient format by using such standard Fortran programs as the Statistical Package for the Social Sciences (SPSS). In this manner all possible two-way contingency tables were constructed using the basic five variables. The study of such contingency tables has a long history (Bishop, Fienberg and Holland 1975) and a variety of analytical procedures is available to the researcher. An evaluation of the relative merits of these procedures lies beyond the scope of this paper; however, the choice of which one to use in a particular case depends on the questions one wishes to have answered.

One question concerns the amount of information which may be lost when the number of attribute categories or sub-categories used to describe the pottery corpus is reduced. Another question addresses itself to the problem of how to refine the dating of this pottery using the large amount of available attribute data.

Following a suggestion by Kruskal (1976), an information measure was used as an analytical tool. Kullback (1959) and Ku et al. (1971) developed methods for analyzing multidimensional contingency tables based on this information measure. It is perhaps appropriate here to echo the remarks of Orloci (1968) who stressed the point that the term information is understood here in a strictly technical sense, conceived as a physical property of events related to probability. In this sense a rare event conveys more information than a common one. Thus, as Orloci states (p.148), "Information as a technical term relates conceptually more closely to surprisal than to either knowledge or informativeness of the ordinary speech."

One property displayed by the Mycenaean pottery data which was tabulated in two-way contingency tables is the appearance of many cells which have zero entries. There are simply no pots of a given type for a given site or time in many instances. These "zero counts" are quite readily handled by the information measure employed below.

Let us take as a specific example the two-way table, Site by Date (Figure 1). The site groups may then be re-ordered from "richest" to "poorest" (the site groups having the least number of pots).

Information is defined as

$$\text{Info} = -p_{ij}\log_e p_{ij}$$

where p_{ij} is the probability that pots have been found on site "i" belonging to time period "j". The logarithms are to the base "e". The data in Figure 1 reproduce the actual number of vessels found at site "i" from time period "j", which is defined as f_{ij}. The f_{ij}'s can be read directly from the table in Figure 1 (row "i", column "j"). For this contingency table p_{ij} can be estimated

$$p_{ij} = \frac{f_{ij}}{N}, \text{ where N is the total number of pots in the table.}$$

A measure of interdependence of rows and columns (Kullback 1959), or, in other words, the amount of information provided by the rows about the columns and vice versa, may be expressed by

$$I(R:C) = \sum_{i=1}^{R} \sum_{j=1}^{C} f_{ij} \log_e \frac{f_{ij}N}{f_{i+}f_{+j}}$$

where f_{i+} = the sum of entries in the i^{th} row

f_{+j} = the sum of entries in the j^{th} column

N = the sum of all entries.

This information measure applied to an archaeological context is mentioned by Peebles (1972) and recommended by Orloci (1969). The importance of values calculated for the I(R:C) statistic can be tested in a way similar to the standard tests used in contingency tables, i.e., by comparison with the chi-squared distribution (Bishop et al. 1975). If such a test shows that I(R:C) is small, or, in other words, insignificant, then rows and columns are nearly independent, and not much is learned about time periods when examining the number of pots at each site or site group. In the Mycenaean case, however, rows (sites) and columns (dates) are far from independent. I(R:C) is not small. Hence, this measure may be used to explore whether the "poorer" sites can be ignored in preliminary analyses in the interest of simplicity and economy. This question becomes even more important when considering pottery types and decorative motives (336 types and 80 motifs) and the quite bulky tables generated by such frequency distributions.

When discarding "poorer" sites from the contingency tables, the information I(R:C) is calculated for different values of R representing the number of sites retained in the Site by Date two-way contingency table. In Figure 2, the percentage loss of information is calculated as rows are progressively discarded (in ascending order from "poor" to "rich") on the ordinate axis. The abscissa shows the corresponding percentage loss of pottery specimen from the total popu-

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES SPSSH - RELEASE [illegible] JOB [illegible]

FILE FURUMARK (CREATION DATE = 08/05/76)

* * * * * C R O S S T A B U L A T I O N O F * * * * *
TABSITE BY TABDATE
PAGE

TABSITE (COUNT)	TABDATE MY I 1	MY IIA 2	MY IIB 3	MY IIIA1 4	MY IIIA2 5	MY IIIB 6	MY IIIC1 7	MY IIIC2 8	ROW TOTAL
1 ARGOLID.CORINTHI	56	62	102	97	169	179	208	3	876 32.9
3 LACONIA	0	6	3	2	4	5	0	0	20 0.8
4 S-W PELOPONNESOS	15	20	1	4	0	2	7	6	55 2.1
6 AETOLIA.ACHARNAN	0	2	0	0	2	1	0	0	5 0.2
7 ATTICA.SALAMIS.M	16	15	24	11	44	50	46	31	237 8.9
8 BOEOT.PHOCIS.LOC	7	23	28	27	33	40	40	6	205 7.7
9 THESS.N.GREECE.M	0	0	11	3	1	6	3	0	24 0.9
10 EUBOEA	0	2	1	6	6	4	0	0	19 0.7
11 CYCLADES	0	1	1	0	4	1	6	0	13 0.5
12 DODECANESE	0	0	5	34	161	113	267	0	580 21.8
13 CRETE	0	0	0	0	2	1	0	1	4 0.2
14 ASIA MINOR	0	1	1	2	3	1	3	0	11 0.4
15 CYPRUS	0	0	3	29	182	286	2	0	502 18.8
16 EGYPT	0	1	0	0	16	11	0	0	28 1.1
17 PALESTINE	0	1	0	1	29	14	0	0	45 1.7
18 SYRIA)	0	0	0	2	6	22	0	0	30 1.1
19 SICILY	0	0	0	0	10	0	0	0	10 0.4
COLUMN TOTAL	94 3.5	134 5.0	177 6.6	218 8.2	672 25.2	736 27.6	588 22.1	45 1.7	2664 100.0

FIGURE 1

lation. For example, if the eight "poorest" site groups are discarded, only five percent of the pottery population is discarded. However, twelve percent of the information is lost about the chronological distribution of Mycenaean pottery, or a factor two and a half times greater. Figure 2 illustrates the need for caution when considering the elimination of sites with few pots for the purpose of simplifying analyses. While this conclusion is not particularly startling to archaeologists, the method employed to support it can be applied to other categories of information about which there is no obvious prior consensus of archaeological opinion. This effect of the loss of information can be improved somewhat by pooling or combining the discarded site groups and re-introducing the pooled data back into the table as one site group, i.e., a catch-all "other" category. In this manner the size of the table can be reduced and more information retained than would be the case when the site groups are completely eliminated. In the example where the eight "poorest" site groups are eliminated, by re-introducing the data as one "other" category, the information loss is reduced from twelve to nine percent.

Further studies are underway to develop models which would describe the basic structure underlying these contingency tables, if such a structure exists. If these efforts should prove fruitful, they may deepen the understanding of the interrelationships among the different variables under consideration, and this information measure seems to offer some help in this investigation.

Richard R. Mertz
Princeton, New Jersey

George J. Levenbach
New Providence, New Jersey

References:

Bishop,Y.M.M., Fienberg S.E., Holland P.W. Discrete Multivariate Analysis (Cambridge, Massachusetts: MIT Press, 1975).

Furumark, Arne The Mycenaean Pottery, Vol. 1 Analysis and Classification: Vol. 2 The Chronology of Mycenaean Pottery (Stockholm, 1941).

Ku, H.H., Varner R.N., Kullback S. "Analysis of Multidimensional Contingency Tables," Journal of the American Statistical Association 66 (1971): 55-64.

Kullback, S. Information Theory and Statistics (New York: John Wiley and Son, 1959).

Kruskal, J.B. (1976), Personal Communication.

Orloci, L. "Information Theory Models for Hierarchic and Non-hierarchic Classifications," in Numerical Taxonomy, A. J. Cole,ed. (London: Academic Press, 1969, pp. 148-163).

Peebles, C.S. "Monothetic Divisive Analysis of the Moundville Burials: An Initial Report," Newsletter of Computer Archaeology 8 (December 1972):1-13.

INFORMATION LOST ABOUT DATES WHEN COMBINING SITES INTO ONE CATEGORY OR WHEN ELIMINATING SITES WHICH HAVE ONLY SMALL QUANTITIES OF POTTERY

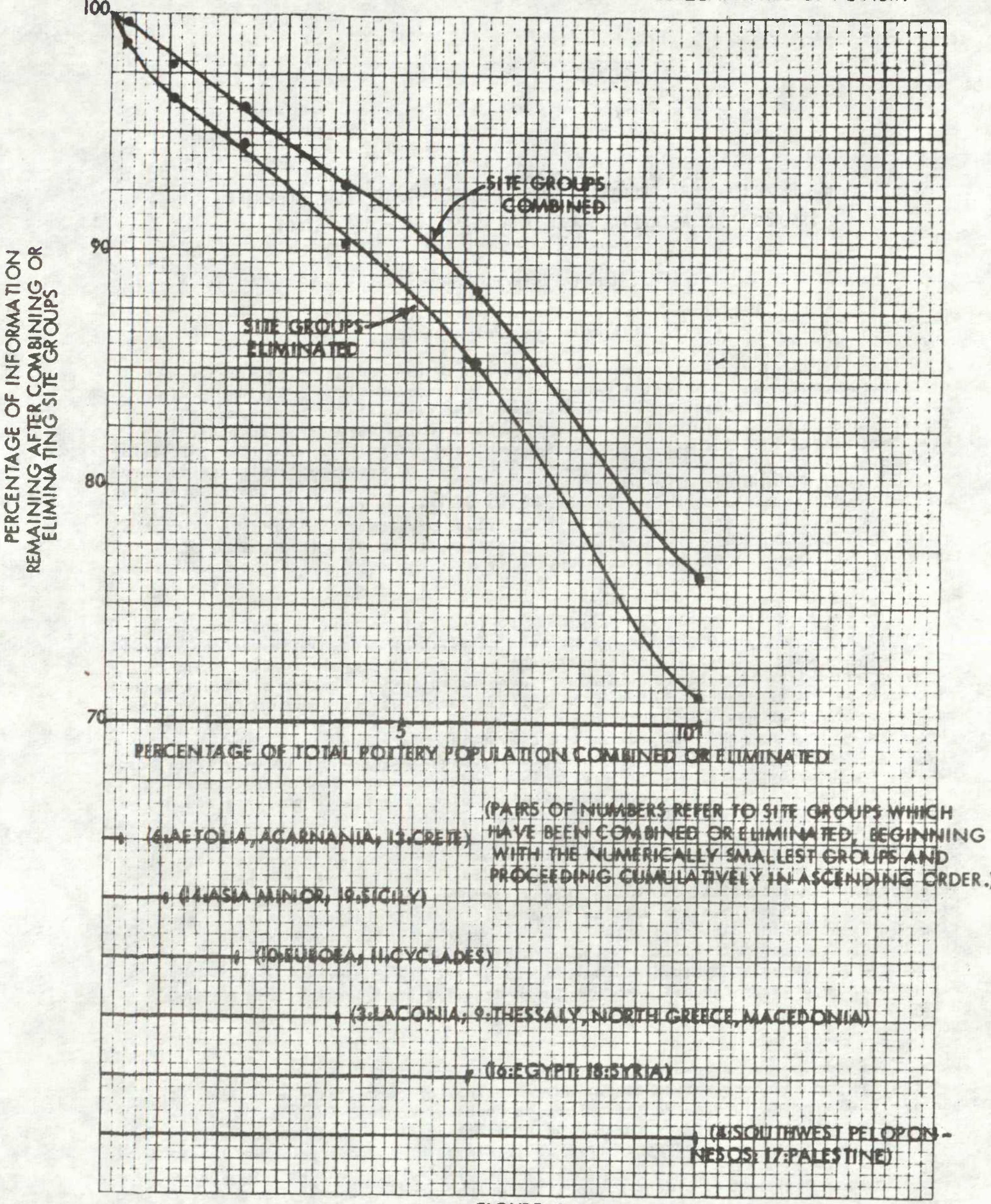

FIGURE 2