ON THE ANALYSIS OF MULTIDIMENSIONAL CONTINGENCY TABLE DATA USING LOG LINEAR MODELS

Geoffrey A. Clark

Department of Anthropology,
Arizona State University, Tempe,
Arizona 85281, U.S.A.

Introduction

Even the most cursory inspection of recent literature makes it apparent that archaeologists are coming to rely more and more heavily upon the use of statistical procedures for data description and analysis (Azoury and Hodson 1973:292-306; Hodson 1970:299-330; McNutt 1973:45-60; Redman 1973:61-79; Weiss 1973). Unfortunately, it seems that statistics are sometimes regarded as substitutes for, rather than adjuncts to rigorous thinking, as scholar after scholar jumps onto this latest of methodological bandwagons. Occasional misapplication is inevitable, however, and does not detract from the tremendous potential inherent in statistical procedures used with rigor to assist traditional methods of problem formulation and solution.

Few would argue, then, that a degree of statistical expertise would be beneficial to most archaeologists. It is unrealistic, however, to expect archaeologists to become statisticians themselves, a time-consuming process beyond the interests of most and the capabilities of many individuals. Nevertheless, the professional should probably take the time to become familiar enough with basic statistical method and theory to be able to evaluate the use of statistical techniques in the literature pertinent to his field. While we decry, and in fact assert the impossibility of the use of statistical methods in a theoretical vacuum, it is apparent that statistical procedures can greatly facilitate problem definition. Whatever the theoretical stance might be which leads to the generation of problems in the broader sense of the term, problems so defined may be described in logically precise ways using inductive statistics, and thus become amenable to analysis through a programme of formal hypothesis formulation and testing.

Below we present one technique which we consider promising. It entails the construction of multidimensional contingency tables which are subsequently analyzed using log linear models (Fienberg 1970:419-433; Goodman 1968:1091-1131; 1969:486-498; 1970:226-256; Muller and Mayhall 1971:149-153). This technique addresses itself to the solution of a fundamental archaeological problem, that of distinguishing important or determinate sources of variation from random variation or "noise". The domain of investigation can be that of artifact, artifact type, feature, site or site aggregate; scale is irrelevant, the structure of the problem is the same at all levels. In the general case, if the total variation measured by variables a, b, c, . . .n is considered to adequately describe variation in a class of data (e.g. an artifact type), it is useful to know which variables are most important, and which contribute little or nothing to the descriptive power of the model employed. The analysis of contingency table data, using log linear models, is one potentially useful approach to the solution of this general kind of problem. We will describe the technique itself, and then illustrate its application with a trivial archaeological example.

Multidimensional Contingency Table Analysis

A contingency table may be defined as a matrix or an array of counts or observations which simultaneously cross-classify objects as belonging to one or more variables, which themselves are present in two or more mutually exclusive states.

A simple two-way contingency table is presented in Table 1. Note that objects are classified according to two multistate variables: Variable 1 is present in three states (C_1, C_2, C_3); Variable 2 is present in four states (C_1, C_2, C_3, C_4). A common approach to this kind of classification problem is to insert raw counts in all cells and convert these data to relative frequencies. This, of course, is done by using the marginal totals as estimators; that is, one can convert to percentages using row totals, column totals or N (the table total) as estimators.

By converting to percentages, one obtains an empirical estimate of the probabilities of obtaining an observation with a given value on Variable 1 and a given value on Variable 2. Counts are thus converted to expressions of probability:

(1) $n_{ij} / N = p_{ij}$

The constraints are those which apply to all probability statements: no given probability can be less than zero (i.e. negative), nor can any given probability exceed one. All probabilities must sum to one.

(2) $p_{ij} \geqslant 0$ p_{ij} 0

(3) $p_{ij} \leqslant 1$; $\sum p_{ij} = \frac{\sum n_{ij}}{N} = 1$

The contingency table format is usually applied to non-metric data; however, it can be used with metrical data (i.e. data which have a continuous underlying distribution) by establishing class intervals and inserting counts in them.

Conventionally, data of this sort are analyzed by using a Chi-Squared Test (Siegal 1956:42-47, 104-111, 175-179). One might ask whether the horizontal distribution is the same for one state within a variable as it is for another, or, generally, how do the relative cell frequencies vary from cell to cell? Are the distributions homogenous or not? Those familiar with X^2, however, will recognise that two constraints limit its usefulness. The first is that expected cell counts must be greater than or equal to some number (usually 5, sometimes 3):

(4) $e_{ij} \geqslant 5$; $e_{ij} \geqslant 3$

Failure to meet this constraint usually leads to the collapsing of the table, which in turn results in lost information. Second, one cannot analyze above 2-way interactions using X^2.

Contingency table analysis allows for expected cell counts to be zero, and permits the examination of higher order (i.e. greater than 2-way) interactions. It also allows for zero raw cell counts, whereas an unmodified X^2 does not. The method is not, however, completely free of constraints. As with X^2, a multinomial distribution is assumed for the data tabulated as a prerequisite for obtaining cell estimates. One consequence of a multinomial distribution is that cells are theoretically independent; thus marginal totals can be used as estimators. A second constraint is that, for obvious reasons, no marginal total used in calculations can contain a zero.

In contingency table analysis, as in X^2, one generates expected counts using the marginal totals derived from a model designed by the investigator. The expected values are the compared with the observed values. The principle difficulty lies in casting investigator-generated hypotheses into explicit statements of relationship between variables. If these hypotheses are properly defined, they can be expressed in the form of a linear equation. It is in the sense of an equation that we use the term "model" here. It is more convenient to express the model in terms of the natural logarithms of the cell probabilities than it is to try to deal with the cell probabilities themselves. For this reason, the model is said to be a "log linear" one.

Those readers familiar with statistical applications will note the similarity between the model described above and the analysis of variance (ANOVA) model. It is useful to consider the case of the ANOVA model in order to explicate and define the terms in the CTAB equation.

Consider the case of a 2-way ANOVA with no replications (the number of replications simply refers to the number of observations taken in each cell). The equation is of the form:

$$(5) \qquad y_{ijn} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijn}$$

where y_{ijn} specifies the row, column and individual within the cell, μ is a constant (the grand mean of the expected cell counts), α_i is the row effect, β_j is the column effect, γ_{ij} is the row/column interaction term, and E_{ijn} is the error term. If n = 1 (i.e. if only one observation is taken per cell), then it is not possible to estimate the interation between the two variables and the formula collapses to:

$$(6) \qquad y_{ij} = \mu + \alpha_i + \beta_j + E_{ij} .$$

In order to cast the Two-Way ANOVA equation into CTAB form, we simply replace the above terms with the relative cell frequencies:

$$(7) \qquad \log p_{ij} = [1] + [A]_i + [B]_j$$

where $\log p_{ij}$ specifies the natural logarithm of an observation identified by its subscript, where $[1]$ is the grand mean of the logs of the expected counts, $[A]_i$ is the main effect due to A_i

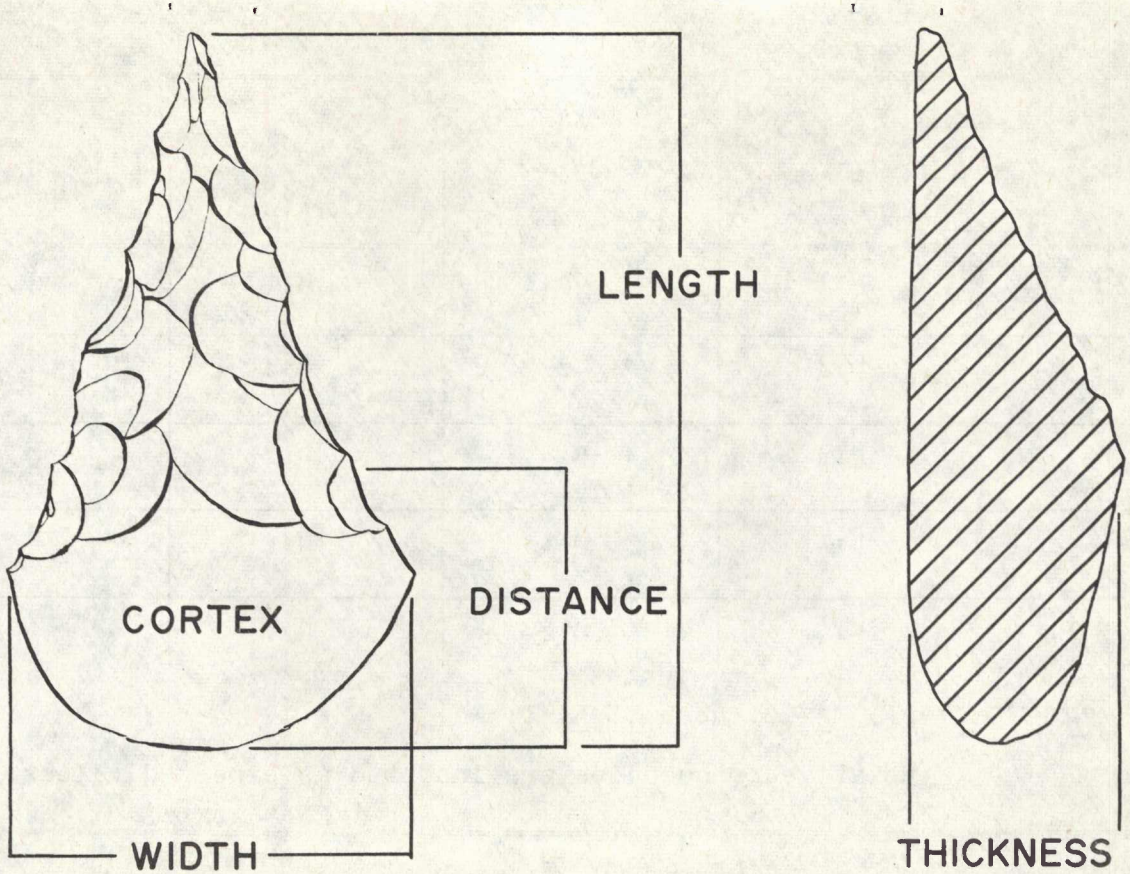


FIGURE 1. VARIABLES MEASURED ON ASTURIAN PICKS. Length (L) is maximum length measured along the axis of the piece; Width (W) is maximum width perpendicular to the axis of the piece; Thickness (Th) is maximum thickness in longitudinal section; Distance (D) measures the maximum extent of unaltered surface (cortex) along the axis of the piece.

VARIABLE 1

VARIABLE 2	C_1	C_2	C_3	
C_1	n_{11}	n_{21}	n_{31}	$n_{.1}$
C_2	n_{12}	n_{22}	n_{32}	$n_{.2}$
C_3	n_{13}	n_{23}	n_{33}	$n_{.3}$
C_4	n_{14}	n_{24}	n_{34}	$n_{.4}$
	$n_{1.}$	$n_{2.}$	$n_{3.}$	N

CONDITIONS: $p_{ij} \geq 0$, no $p_{ij} \geq 1$; $\Sigma n_{ij} / N = 1$.

Relative frequencies are inserted as expected values.

TABLE 1. A CONTINGENCY TABLE SIMULTANEOUSLY CROSS-CLASSIFYING TWO MULTISTATE VARIABLES. Variable 1 is present in three states (C_1,C_2,C_3); Variable 2 is present in four states (C_1,C_2,C_3,C_4).

and $[B]_j$ is the main effect due to B_j. The main effects refer to a specified set of marginal totals selected by the model to be fit. The method will generate expected cell values, on the basis of this particular subset of marginal totals. These expected values are referred to as maximum likelihood estimates; they express the most probable values for the observed cell counts to take on IF THE MODEL CHOSEN IS CORRECT.

The technique then compares the maximum likelihood estimates with the original observed cell counts. If the main effects fit by themselves, then it can be assumed that the interaction terms are negligible (i.e. they approximate zero). If the expected values generated by the model do not agree with the observed values, then a non-zero interaction exists. The marginals used to generate the expected values are the highest order interactions in the model. The importance of zero marginals becomes clear: if any marginal total sums to zero, then no estimates can be obtained from it. Zero marginals are usually eliminated by adding a small constant (e.g. .01) to all tabulated values.

Given the similarity of this method to analysis of variance, it is pertinent to ask what advantages CTAB might have over ANOVA. The main reason contingency table analysis is to be preferred is that it is not characterized by the strong underlying assumption of normality which is a feature of analysis of variance. Also, zero cell counts are possible in CTAB analysis; they must be corrected for in ANOVA.

Model Formulation

We turn now to the question of model formulation. It is obvious that given even a few primary variables, a comparatively large number of models can be generated; 2^n models will result for n primary variables. Two major approaches have been developed to generate and evaluate models of the form described above. They can be labelled the Fienberg and the Goodman approaches, although those authors are not unique in their contributions to the problem.

The Fienberg Approach

Stephen Fienberg (1970:419-433), a statistician at the University of Chicago, has developed a method which takes a series of models, each one of which represents a set of explicit hypotheses about the data, orders these models into a hierarchy and evaluates that hierarchy on the criteria of adequacy and parsimony. Hierarchical models are models (in this case equations) ordered from simple to complex, such that any given model contains all of the terms in the model which precedes it. In the context of a contingency table analysis, this means that if an interaction term (AB) occurs in the model, then the primary variables (A) and (B) must also be included. It might be the case that the investigator regards the primary variables (A) and (B) by themselves as meaningless; nevertheless, they

must be included in the equation.

The Fienberg approach has the advantage of greater precision, but assumes considerable forehand knowledge of the behaviour of the data. Considerable thought about the hypotheses to be tested is a prerequisite, but the technique is more "elegant" in the mathematical usage of the word. It has the disadvantage that it might not always prove to be adequate if the behaviour of the data is completely unknown, or if its behaviour is "masked" by unforeseen and complex interactions.

The Goodman Approach

The second approach, outlined in a series of papers by Leo Goodman (1968:1091-1131; 1969:486-498; 1970:226-256), fits the most complex (most complete possible) model to the data, and then tests whether the effects due to each term are zero or not. In this way the terms in the model are successively reduced until all zero terms are eliminated, resulting in the simplest, adequate model.

The Goodman approach has the advantage that it cannot fail to produce a model which adequately describes the pattern of variation in the data. The variables isolated, however, might be so complex that they defy interpretation. No previous knowledge of the data is required under the Goodman approach; there is no necessity to formulate explicit hypotheses. By comparison with Fienberg's approach, this method is "sloppy" in the sense that a lot of extraneous information goes into the construction of the "most complete" model. In either case, the final objective is to isolate the simplest and most comprehensive model.

Decision Making Criteria

Given that a number of models will be generated by the analysis, one must face the problem of how these models are to be compared if the isolation of a single "best" model is the objective.

The obvious first step is to determine whether a model "fits" the data or not; that is, whether the expected cell counts are good predictors of the observed cell counts. It will probably be the case that a number of models "fit" the data in the sense defined above; the second step is to make a choice among them. The only constraint for comparison is that the models be of a hierarchical nature (i.e. ordered from simple to complex); if they are not, the tests used to compare them cannot assume independence.

The two models most frequently used to compare models are X^2 and the log likelihood ratio ($\log \lambda$). Chi-squared tests are widely known and used; they require no further comment. The log likelihood ratio also makes use of the X^2 distribution. If λ = the likelihood ratio, the expression

(8) $\quad -2 \log \lambda$ approximates the X^2 distribution.

The log likelihood ratio is obtained by taking the log of each quotient (observed / expected) cellwise, summing the logs, and multiplying by two:

$$(9) \qquad -2 \log \lambda = 2 \sum (\log O/E).$$

The values for the X^2 distribution are well tabulated.

Although both X^2 and $\log \lambda$ are suitable methods for testing the difference between models, $\log \lambda$ has the advantage that it can be partitioned into independent parts such that each partition is an independent test of a particular model. Chi-squared cannot be so partitioned. $\log \lambda$ is also more stable for small values (≤ 5) than is X^2.

The steps discussed so far are simple but tedious if done by hand. There is, however, a computer program (CTAB) in the SNAP series (University of Chicago) which provides output specifying cell estimates, log likelihood ratio and degrees of freedom fit for each model tested. All that it is necessary to do, is to draw up a table showing the log likelihood ratios and degrees of freedom fit for each model. Since the models are hierarchical, one can use these statistics to test differences between them. Evaluation proceeds pairwise from the most complex model to the simplest. Two stopping criteria are employed: (1) when Model X adequately describes the data and Model Y does not, choose Model X; (2) when Model X and Model Y both describe the data, and there is a statistically significant difference between them choose Model X. Because of the hierarchy and the evaluation procedure used, Model X will always be the simpler of the two.

An Archaeological Example

An illustration of the method using a concrete archaeological example is presented below. Data come from an assemblage known as the Asturian of Cantabria (Vega del Sella 1923; Clark 1971a; 1971b), found in the provinces of Asturias and Santander, on the north coast of Spain. Sites consist of semi-brecciated midden deposits located in cave mouths along the Cantabrian littoral. Large, crude quartzite tools form an important component of the lithic industry. The assemblage dates to the early Holocene (8,900-6,000 BP)(Clark 1971b:1245-1257).

The sample selected for analysis consisted of 92 pointed, uni-facial quartzite core tools called "Asturian Picks". These implements are the so-called "guide fossil" for the industry. Each pick was classified by site and by a series of four rather trivial dimensions: length (L), width (W), thickness (Th) and distance (D)(Fig. 1). Dimensions were trivial because little confidence can be placed in provenience data, owing to inadequate cataloging procedures. The high probability of mixed collections did not justify more elaborate recording of attribute data. Nevertheless, the data selected are adequate to illustrate the method outlined above; however, no attempt will be made to draw culturally relevant conclusions from the analysis.

The five variables used are listed in Table 2; the variable "site" was present in six states and each of the four dimensions was subdivided into "large" (ℓ) and "small" (σ). Subdivisions

within dimensions are, in this case, arbitrary. All dimensions were plotted and were found to have unimodal distributions; consequently, no obvious criteria for subdivision was available. The median was selected as the criterion for dividing "large" from "small". The median was employed for this purpose because it is a better measure of central tendency than the mean; the latter is influenced by outliers. The result is a 5-way contingency table, formed by a 6 x 2 x 2 x 2 x 2 matrix and consisting of a total of 96 cells.

Table 3 shows the actual contingency table. Note the high frequency of zeros and low cell values, both features which would have made X^2 or conventional ANOVA difficult or impossible. Fig. 2 is simply an attempt to depict the matrix more accurately; it is, of course, impossible to draw a five-dimensional space.

The Fienberg Approach

We sought first to apply the Fienberg approach to the problem. A non-parametric test called the Kruskal-Wallis H Test (Wallis and Roberts 1967:599-601; Siegal 1956:185-193) was applied to the data as a preliminary step in order to derive the series of explicit models demanded by Fienberg's method. The Kruskal-Wallis H test is a simplified 1-way analysis of variance; it does not assume a normal distribution. The test simply evaluates whether or not the medians of k samples are derived from populations having the same or similar underlying distributions. The formula:

$$(10) \qquad H = \frac{12}{N\,(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(N+1)$$

where N is the total number of observations in all samples, k is the number of samples (in this case, sites), n_i is the number of observations in a given sample and R_i^2 is the sum of the ranks squared for any given sample.

The results of the test indicated that, with respect to site, there are differences in the length and distance measurements of the picks, but none with respect to width and thickness. The implication is that the effects due to width and thickness are not important by themselves; therefore, they were not included in the hierarchy of models formulated on the basis of the Kruskal-Wallis test. It is worth commenting, parenthetically, that the Kruskal-Wallis test evaluates only main effects; in fact it will be demonstrated below that it is the interactions, rather than the main effects, which constitute the important variables.

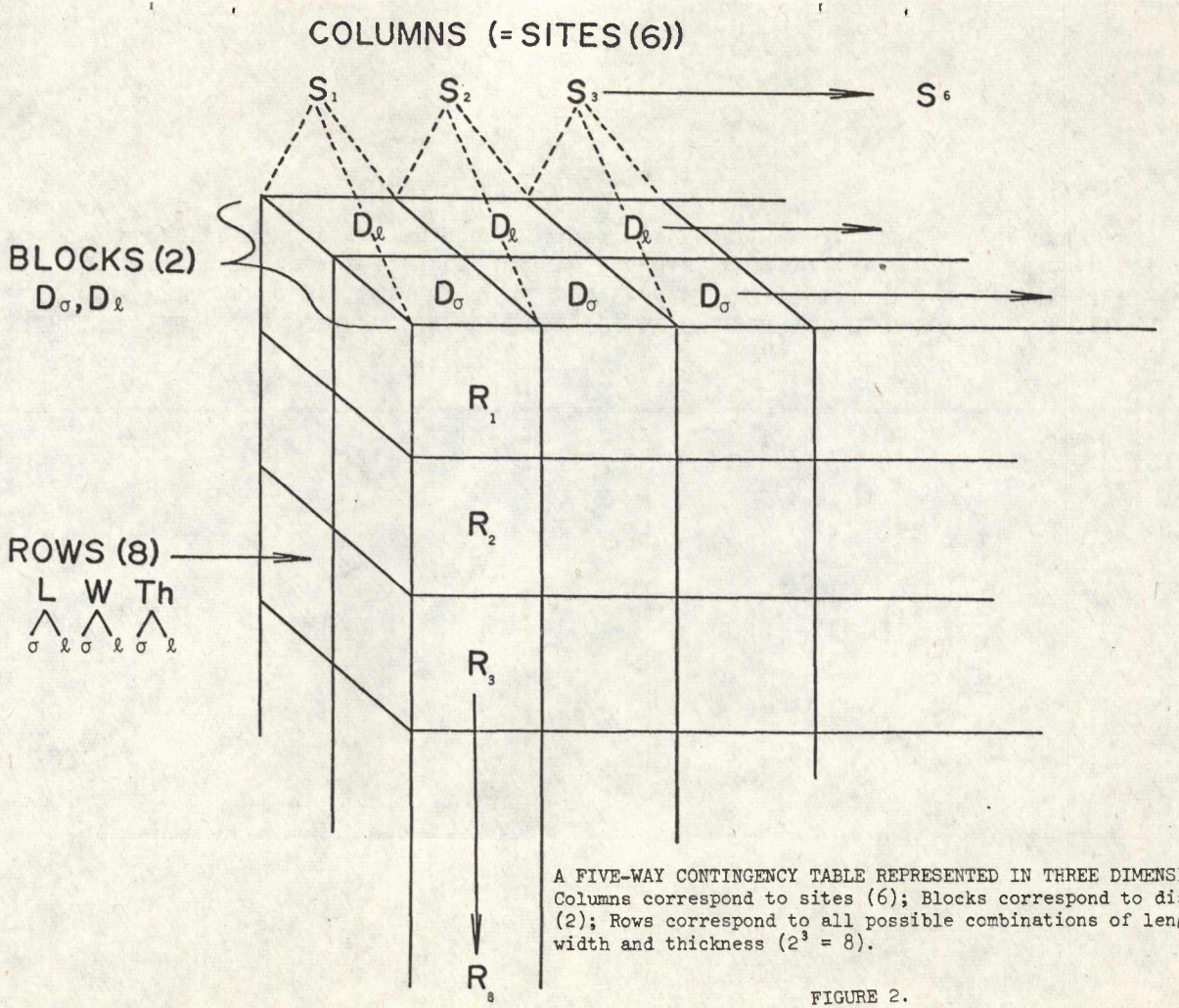
The series of hierarchically ordered log-linear models developed using the Kruskal-Wallis test as a basis are presented in Table 4. The CTAB program generates log-likelihood ratios and degrees of freedom fit for each model run. Evaluation simply entails consultation of a X^2 table at some predetermined level of significance (in this case .01 and .05 were both used).

SITES:	LA RIERA LLEDIAS		ARNERO PENICIAL		COBERIZAS FONFRIA
LENGTH:	LARGE > 8.0 SMALL < 8.0		WIDTH:	LARGE > 5.6 SMALL < 5.6	
DISTANCE:	LARGE > 4.1 SMALL < 4.1		THICKNESS:	LARGE > 3.1 SMALL < 3.1	

TABLE 2. FIVE MULTISTATE VARIABLES RECORDED ON ASTURIAN PICKS: SITE PROVENIENCE (La Riera, Lledias, Penicial, Arnero, Coberizas, Fonfria); LENGTH (large(ℓ), small(σ)); WIDTH (large, small); THICKNESS (large, small); DISTANCE (large, small). Grand medians were used to distinguish large (ℓ) from small (σ) with respect to the continuously distributed metrical variables.

SITES:	LA RIERA n=36		LLEDIAS n=11		PENICIAL n=11		ARNERO n=13		COBERIZAS n=10		FONFRIA n=11	
	D_σ	D_ℓ	D_σ	D_ℓ	D_σ	D_ℓ	D_σ	D_ℓ	D_σ	D_ℓ	D_σ	D_ℓ
L_σ W_σ Th_σ	3	3	7	-	-	-	-	-	2	-	1	2
L_σ W_σ Th_ℓ	3	1	-	-	1	1	3	4	2	-	1	-
L_σ W_ℓ Th_σ	3	2	-	1	-	-	-	-	1	-	1	1
L_σ W_ℓ Th_ℓ	2	-	-	-	-	-	1	3	1	2	1	-
L_ℓ W_σ Th_σ	-	2	-	-	-	-	-	-	-	-	-	2
L_ℓ W_σ Th_ℓ	3	-	-	-	-	4	-	1	1	-	-	-
L_ℓ W_ℓ Th_σ	2	4	-	1	-	-	-	-	-	1	-	2
L_ℓ W_ℓ Th_ℓ	4	4	2	-	4	1	1	-	-	-	-	-

TABLE 3. A CONTINGENCY TABLE SIMULTANEOUSLY CROSS-CLASSIFYING FIVE MULTISTATE VARIABLES RECORDED ON ASTURIAN PICKS. The matrix is of the form (6 x 2 x 2 x 2 x 2); it contains a total of 96 cells.



A FIVE-WAY CONTINGENCY TABLE REPRESENTED IN THREE DIMENSIONS. Columns correspond to sites (6); Blocks correspond to distance (2); Rows correspond to all possible combinations of length, width and thickness ($2^3 = 8$).

FIGURE 2.

MODEL 1 : $\log p_{ijklm} = [1] + [S]_i$

MODEL 2 : $\log p_{ijklm} = [1] + [S]_i + [L]_j + [SL]_{ij}$

MODEL 3 : $\log p_{ijklm} = [1] + [S]_i + [L]_j + [SL]_{ij} + [D]_k + [SD]_{ik}$

MODEL 4 : $\log p_{ijklm} = [1] + [S]_i + [L]_j + [SL]_{ij} + [D]_k + [SD]_{ik} + [LD]_{jk}$

MODEL 5 : $\log p_{ijklm} = [1] + [S]_i + [L]_j + [SL]_{ij} + [D]_k + [SD]_{ik} + [LD]_{jk} + [SLD]_{ijk}$

TABLE 4. AN EXAMPLE OF THE FIENBERG APPROACH: HIERARCHICALLY ORDERED LOG LINEAR MODELS BASED ON THE ASTURIAN PICK DATA. As a preliminary step, the Kruskal-Wallis H test was used to isolate potentially important variables; site, length and distance were selected. The main effects so isolated were subsequently combined in linear equations to include all 2- and 3-way interactions

MODEL	LOG-LIKELIHOOD RATIO	DEGREES OF FREEDOM (FIT)	CHI-SQUARED α=.05	α=.01
MODEL 1	84.11	6	S	S
DIFFERENCE 1/2	10.56	6	NS	NS
MODEL 2	73.55	12	S	S
DIFFERENCE 2/3	4.86	6	NS	NS
MODEL 3	68.69	18	S	S
DIFFERENCE 3/4	2.15	1	NS	NS
MODEL 4	66.54	19	S	S
DIFFERENCE 4/5	1.75	5	NS	NS
MODEL 5	64.79	24	S	S

TABLE 5. AN EVALUATION OF TABLE 4: THE LOG-LIKELIHOOD RATIO AS A DECISION MAKING CRITERION. Each model defined is evaluated for descriptive adequacy using the χ^2 distributed log-likelihood ratio; differences between models are also assessed. S indicates a significant difference between observed and expected values; a model so described does not fit the data. NS indicates a non-significant difference between expected and observed values; a model so described does in fact adequately describe the pattern of variation in the data. Results: only one model is represented; differences between models are not significant. No model adequately describes the data.

$$\text{MODEL 1 : } \log p_{ijklm} = [1] + [S]_i + [L]_j + [D]_k + [W]_l + [T]_m + [SL]_{ij} +$$
$$[SD]_{ik} + [SW]_{il} + [ST]_{im} + [LD]_{jk} + [LW]_{jl} + [LT]_{jm} +$$
$$[DW]_{kl} + [DT]_{km} + [WT]_{lm} + [SLD]_{ijk} + [SLW]_{ijl} +$$
$$[SLT]_{ijm} + [SDW]_{ikl} + [SDT]_{ikm} + [SWT]_{ilm} + [LDW]_{jkl} +$$
$$[LDT]_{jkm} + [LWT]_{jlm} + [DWT]_{klm} + [SLDW]_{ijkl} +$$
$$[SLDT]_{ijkm} + [SLWT]_{ijlm} + [SDWT]_{iklm} + [LDWT]_{jklm}$$

TABLE 6. AN EXAMPLE OF THE GOODMAN APPROACH. The complete model is given, incorporating 31 terms. There are five main effects, ten 2-way interactions, ten 3-way interactions and five 4-way interactions.

MODEL	LOG-LIKELIHOOD RATIO	DEGREES OF FREEDOM (FIT)	CHI SQUARED α=.05	α=.01
MODEL 1	2.19	6	NS	NS

TABLE 7. AN EVALUATION OF TABLE 6: THE LOG-LIKELIHOOD RATIO AS A DECISION MAKING CRITERION. The complete model adequately describes the pattern of variation in the data.

VARIABLES ELIMINATED BY INSPECTION OF U-ESTIMATES:

LENGTH	TL	SD	SDL	DTL	SDTW
WIDTH	DT	DL	SWL	DWL	
THICKNESS	SW	DW	WL	TWL	
DISTANCE	TW	STW	DTW	LWTD	

VARIABLES POSSIBLY IMPORTANT:

SL	SDW	SDTL
SDT	STL	STWL

VARIABLES DEFINITELY IMPORTANT:

SITE	ST	SDWL

TABLE 8. RESULTS OF THE INSPECTION OF U-ESTIMATES. U-values less than .20 were eliminated from further consideration; U-values greater than .20 but less than .50 were regarded as possibly important; U-values greater than or equal to .50 were considered definitely important in data description (NB: see Footnote 1).

MODEL A : $\log p_{ijklm} = [1] + \underline{[S]}$

MODEL B : $\log p_{ijklm} = [1] + [S] + [T] + \underline{[ST]}$

MODEL C : $\log p_{ijklm} = [1] + [S] + [D] + [W] + [L] + \underline{[SDWL]}$

MODEL D : $\log p_{ijklm} = [1] + [S] + [T] + \underline{[ST]} + [D] + [W] + [L] + \underline{[SDWL]}$

MODEL E : $\log p_{ijklm} = [1] + [S] + [D] + [W] + [L] + \underline{[SDWL]} + [T] + \underline{[STWL]}$

MODEL F : $\log p_{ijklm} = [1] + [S] + [D] + [W] + [L] + \underline{[SDWL]} + [T] + \underline{[SDTL]}$

MODEL G : $\log p_{ijklm} = [1] + [S] + [D] + [W] + [L] + \underline{[SDWL]} + [T] + \underline{[SDTL]} + \underline{[STWL]}$

TABLE 9. THE REDUCED TERM MODELS. Note that the models incorporate only those terms regarded as important by the inspection of the U-estimates. Subscripts are eliminated for clarity.

Models which are SIGNIFICANTLY DIFFERENT (S) for specified α are those which DO NOT fit the data; these are eliminated. Models which are NOT SIGNIFICANTLY DIFFERENT (NS) adequately describe the pattern of variation in the data; these are retained and further evaluated. Differences between models retained are also tested by the log-likelihood ratio.

Table 5 presents the evaluation of the models formulated on the basis of the Kruskal-Wallis H test. The result is clearcut; no model adequately describes the observed data. It is possible, then, to eliminate Models 1 - 5 from further consideration; terms expressing main effects are not adequate in themselves to explain or describe variation in the data.

The Goodman Approach

Given the failure of one set of explicit hypotheses, the investigator has the option of defining other sets, on the basis of different criteria, or resorting to the Goodman approach to isolate important variables. As noted, the Goodman approach defines a single model incorporating all main effects and all possible interaction terms (Table 6). In this case, the model contains a total of 31 terms, including the main effects, 10 2-way interactions, 10 3-way interactions and 5 4-way interactions. As expected, the model fits the data in that it adequately describes them (Table 7); however, no distinction can be made between those variables which are important and those which are not. The results are, at this stage, uninterpretable. As in analysis of variance, however, relative estimates of the effects in the model can be obtained.

For each model tested, the CTAB output produces statistics called estimated U-values. These assess the influence of each term in the equation against the total descriptive power of the equation. Variables with low U-values ($\leq .20$) probably do not play an important role in data description and may be eliminated. Models can be made ever more explicit by successive runs, systematically eliminating terms with low U-values.

An examination of U-values in the most complete model permitted the elimination of 21 terms (Table 8). U-values less than .20 were regarded as insignificant; associated terms were consequently deleted. * The result is immense simplification; only three terms are regarded as definitely important variables (U-values $\geq .50$); six terms are possibly important (U-values $> .20$ but $< .50$).

The final step is to construct a set of models using only those terms regarded as important variables. These models are

---

* It should be noted that this elimination procedure is a practical and useful, but essentially impressionistic approach to the deletion of unimportant terms. Goodman (1969:486-498) advocates a more rigorous evaluation procedure; each term is tested to determine wether a non-zero interaction exists. Only zero interactions are eliminated.

presented in Table 9. Inspection of the models reveals three important points. Note first that no 2- and 3-way interaction terms appear to be included. These terms are actually included in any model which contains a 4-way interaction term; because of the program format used, it is not necessary to specify them. Second, note that all of the terms regarded as important are underlined. Other terms are incorporated into the models because of a constraint of contingency table analysis mentioned earlier: all interactions must have their terms defined (i.e. if (AB) is in the equation, (A) and (B) must also be specified). Finally, note that the models are not entirely hierarchical. Inspection reveals that Model A is a subset of B and C; B and C are subsets of D (but not of each other); D is a subset of E and F; and E and F are subsets of G (but not of each other).

The models are ordered from simple to complex, and the partial hierarchy is represented graphically in Fig. 3. Employing the two stopping criteria defined above, evaluation proceeds from the most complex model (G) to the simplest model (A). Models D, E, F and G all adequately describe the data, moreover, there are no statistically significant differences between them. Model C also describes the data, but is significantly different from Model D. While adequate in terms of the arbitrarily selected levels of significance, it explains the data less completely than does Model D. Models A and B do not adequately describe the data; they can be eliminated from further consideration.

The first stopping criterion (X describes the data, Y does not) is applied to select Model D over Model B. The second stopping criterion (X and Y describe the data, but there is a significant difference between them) results in the selection of Model D over Model C. The application of the stopping criteria both result in the selection of Model D. Model D consists of the 2-way interaction (ST) and the 4-way interaction (SDWL).

The conclusion is that these two variables are the most important in describing variation among samples of picks from Asturian sites (at least insofar as that variation is measured by the trivial variables selected for this example). One might speculate, however, that the variables (T) and (DWL) are behaving in different ways with respect to the variable (S). It might be argued that the (ST) interaction still reflects the original dimension of the flattened, oval cobbles on which the picks are manufactured. Quartzite cobbles occur in the stream beds and estuaries along which Asturian sites are distributed. If raw material adjacent to the site was utilized, one would expect sites and thicknesses to vary together. The difficulty with this is that the cobbles in a stream gravel vary greatly in size according to extremely localized conditions (e.g. gradient). Therefore, one would expect a range of cobbles of differing sizes to be available in the immediate vicinity of a site. However, if thickness was important to the site occupants, and if they were selecting cobbles of certain dimensions, this selection might be reflected in the (ST) interaction. It seems probable that the original thicknesses of the cobbles selected were not altered much by the manufacturing process. The (SDWL) interaction, on the other hand, might reflect variation due to the manufacturing process. Distance, width and length measure

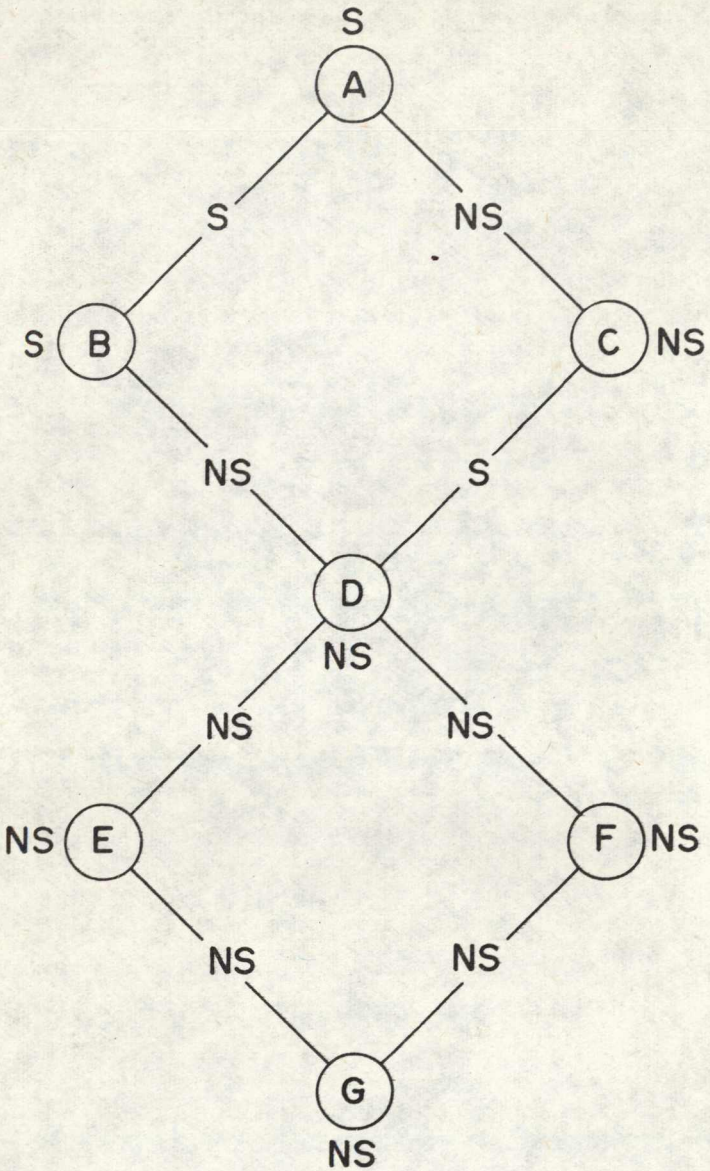


FIGURE 3. THE REDUCED TERM MODELS ORDERED FROM SIMPLE (A) TO COMPLEX (G). Note that the graphical presentation reflects the partial hierarchy described in the text. Evaluation proceeds from bottom to top; $\alpha = .01$. Model D is selected on the criteria of adequacy and parsimony.

the extent to which the original cobble was modified to conform to a culturally-defined ideal. One would expect these variables to be correlated with sites as the manufacturing process for picks was essentially the same across all Asturian sites (Clark 1971a:268,269). In short, the (ST) interaction might reflect human selection for a natural dimension; the (SDWL) interaction might reflect the imposition of technological attributes on a natural object. Taken together, the two interactions adequately describe variation among the Asturian picks used in this example. Whether these same interactions would be isolated using different samples remains to be determined.

Summary

A method for analyzing data cast into contingency table format is presented. A series of models in the form of linear equations ordered in a hierarchy express relationships suspected among the variables selected for evaluation. Marginal totals corresponding to terms in the models are used to generate expected cell values; expected and observed cell values are compared using a X^2 distributed statistic called the log likelihood ratio ($\log \lambda$). Models are evaluated on the criteria of adequacy and parsimony; a "best" model is isolated. The "best" model is the simplest model which adequately describes the pattern of variation in the data. A simple example using archaeological data is presented to illustrate the approach.

Acknowledgements

I wish to express my sincere gratitude to Mr. T.P. Muller, the statistical consultant for the Department of Anthropology, University of Chicago, for invaluable assistance rendered during the planning and execution phases of the various techniques described in this paper. Without his help, and direction to pertinent source material, this paper would probably never have been written. The author, however, is solely responsible for overall content and for any factual or conceptual errors which the manuscript might contain. I also acknowledge the assistance of various members of the Department of Anthropology, Arizona State University, who read and criticized the manuscript at various stages in its development. Especially helpful were P.R. Fish, L.D. Smith, B. Domeier, B.L. Stark and J.D. Cadien.

REFERENCES CITED

Azoury, I. & Hodson, F.R. 1973 Comparing Paleolithic Assemblages: K'sar Akil, a case study. WORLD ARCHAEOLOGY 4:292-306.

Clark, G.A. 1971a- The Asturian of Cantabria: A re-evaluation. Unpublished Ph.D. Dissertation, Department of Anthropology, University of Chicago.

1971b- The Asturian of Cantabria: Subsistence Base and the Evidence for Post-Pleistocene Climatic Shifts. AMERICAN ANTHROPOLOGIST 73 (5): 1244-1257.

Fienberg, S. 1970 The Analysis of Multidimensional Contingency Tables. ECOLOGY 51: 419-433.

Goodman, L.A. 1968 The Analysis of Cross-classified Data: Independence, Quasi-Independence and Interactions in contingency tables with or without missing entries. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 63:1091-1131.

1969 On Partitioning X^2 and Detecting Partial Association in Three-way contingency tables. JOURNAL OF THE ROYAL STATISTICAL SOCIETY 31:486-498.

1970 The Multivariate Analysis of Quantitative Data: Interactions among multiple classifications. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 65: 226-256.

Hodson, F.R. 1970 Cluster Analysis and Archaeology: some New Developments and Applications. WORLD ARCHAEOLOGY 1: 299-320.

McNutt, C.H. 1973 On the Methodological Validity of Frequency Seriation. AMERICAN ANTIQUITY 38:45-60.

Muller, T.P. & Mayhall, J. 1971 Analysis of Contingency Table Data on Torus mandibularis using a Log Linear Model. AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 34:149-153.

Redman, C.L. 1973 Multistage Fieldwork and Analytical Techniques. AMERICAN ANTIQUITY 38:61-79.

Siegal, S. 1956 NON-PARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES. International Student Edition, pp. 42-47, 104-111, 175-179. McGraw Hill and Kogakusha. New York and Tokyo.

Vega del Sella, el Conde de la. 1923 El Asturiense: Nueva Industria pre-neolitica. COMISION DE INVESTIGACIONES PALEONTOLOGICAS Y PREHISTORICAS, MEMORIA NUM. 32 (Serie prehistorica Num. 27). Museo Nacional de Ciencias Naturales, Madrid.

Wallis. W.A. & Roberts, H.V. 1967 STATISTICS: A NEW APPROACH. The Free Press, New York.

Weiss, K.A. 1973 Demographic Models for Anthropology. MEMOIRS OF THE SOCIETY FOR AMERICAN ARCHAEOLOGY No. 27. Washington.