

LINEAR FLOW GRAPHS FROM CONTINGENCY TABLES: A CONDITIONAL PROBABILITY
APPROACH TO LAZARSFELDIAN CAUSAL ANALYSIS*

James R. Beniger
Department of Sociology
Princeton University

Because is in your mind.

--Screamin' Jay Hawkins
(1968), a painting by
Karl Wirsum, American

In contingency table analysis, the concept of "because" -- which may well be only a state of mind--must be introduced by means of an asymmetric model. Unlike symmetric models based on odds ratios, which have already captured the attention of social scientists and applied statisticians (see Goodman 1970; 1972; an encyclopedic overview is provided by Bishop, Fienberg and Holland 1975), asymmetric models--based on proportions rather than odds ratios--have been slower to gain acceptance (but see Goodman 1963; Coleman 1970; Davis 1975b). Nevertheless, asymmetric models seem to hold considerable promise for the type of causal and dynamic contingency table analysis developed by Paul Lazarsfeld and his Columbia colleagues in the 1950s (e.g., Kendall and Lazarsfeld 1950), and still predominant in much of survey, communications and market research.

A particularly promising approach to asymmetric analysis is that of linear flow graphing; this application was first suggested by Huggins and Entwisle (1968). Stinchcombe (1968) introduced the technique into systematic social theory construction, and Heise (1975) employed it to develop major principles of path analysis. Davis (1975a) is a comprehensive treatment of d-system flow graphing.

The purpose of this paper is to motivate the asymmetric analysis of contingency tables using proportions (differences in row and column percentages) in terms of conditional probability. This approach affords a natural causal interpretation, in the sense of changes in future probabilities, for linear flowgraph analysis of nominal or categorical variables. Extensions of the concept of causality to logical and set theoretic notions is also suggested.

Section 1 introduces the notion of partial or contributing cause, for which a measure (the coefficient d_{BA}) is proposed in Section 2. This measure is extended to contingency tables in Section 3, and to causal flow graphs in Section 4. Examples using the data of J.A. Davis (1975a) on region, education and racial tolerance from the NORC General Social Survey are given for the two-event case in Sections 5 and 6, and the three-event case in Sections 7 and 8.

1. Introduction

Consider two events, A and B, such that A can be assumed, for extra-mathematical reasons, to

*This paper was written while the author was a graduate student in the Departments of Sociology and Statistics, University of California, Berkeley. The research was funded, in part, by fellowship 1 F31 DA 05082-01, awarded by the National Institute on Drug Abuse, DHEW.

be a contributing or partial cause (i.e., neither a necessary nor sufficient cause) of B (e.g., A is temporally prior to B). Then the probability of B, given that A has occurred, is greater than it is when A has not occurred, i.e.,

$$P(B/A) > P(B/A^*) \quad (1)$$

This is a necessary but not sufficient condition for A to be a cause of B, controlling for all confounding effects of other events. (The case in which A has a negative or dampening effect on B, i.e., where $P(B/A) < P(B/A^*)$, may be of equal substantive interest; this case can be treated as equivalent to event (1) by reversing the definitions of A and A^*). Note two special cases of (1): when

$$P(B/A^*) = 0, \quad (2)$$

A is said to be a necessary cause of B, and when

$$P(B^*/A) = 0, \quad (3)$$

A is said to be a sufficient cause of B; A is said to be a necessary and sufficient cause of B whenever the intersection of events (2) and (3) obtains.

2. Measuring Partial Causes

Given that A is a partial cause of B, it is often of substantive interest to measure the degree or strength of the relationship between A and B. This task might be seen as one of decomposing the probability that B will occur, $P(B)$, into "explained" (by the occurrence of A) and "unexplained" probabilities. The unexplained probability, call it d_{BA} (as in regression and path notations, the first subscript (B) denotes the dependent event or effect, the second subscript (A) denotes the independent event or cause), will be a function of $P(A)$. What is the expression for d_{BA} ?

From the definition of conditional probability,

$$P(B/A) = P(B \cap A) / P(A), \quad (4)$$

and the fact that

$$P(B) = P(B \cap A^*) + P(B \cap A), \quad (5)$$

it follows that

$$P(B) = P(B/A^*)P(A^*) + P(B/A)P(A). \quad (6)$$

Substituting $1-P(A)$ for $P(A^*)$, equation (6) becomes

$$P(B) = P(B/A^*) + P(A)[P(B/A) - P(B/A^*)]. \quad (7)$$

Equation (7) is in the desired form, namely,

$$P(B) = \underbrace{P(B/A^*)}_{\text{unexplained}} + \underbrace{d_{BA}P(A)}_{\text{explained}}, \quad (8)$$

where

$$d_{BA} = P(B/A) - P(B/A^*). \quad (9)$$

The coefficient d_{BA} has several desirable properties as a measure of the degree of $P(B)$ "explained" by A . When A and B are independent, i.e., when $P(B) = P(B/A) = P(B/A^*)$, then $d_{BA} = 0$. When A is a necessary cause of B (i.e., when (2) holds), the "unexplained" term $P(B/A^*)$ equals 0, and d_{BA} becomes $P(B/A)$, i.e., the entire $P(B)$ is explained by A .

3. Contingency Tables

Readers familiar with contingency table analysis will recognize d_{BA} from equation (9) as a difference in proportions in a two-by-two table. Consider the table

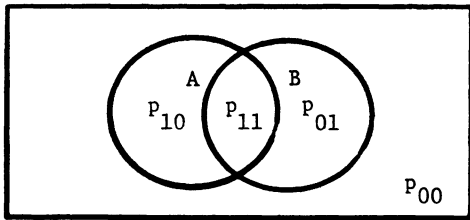
		0	B	1	
	0	P ₀₀	P ₀₁	P _{0.}	
A					
	1	P ₁₀	P ₁₁	P _{1.}	
				P _{.0}	P _{.1}

Here the marginal probabilities are the probabilities of individual events,

$$\begin{aligned} P_{0.} &= P(A^*) & P_{.0} &= P(B^*) \\ P_{1.} &= P(A) & P_{.1} &= P(B) \end{aligned}$$

The cell probabilities are the probabilities of the four possible intersections of A and B ,

$$\begin{aligned} P_{00} &= P(A^* \cap B^*) \\ P_{10} &= P(A \cap B^*) \\ P_{01} &= P(A^* \cap B) \\ P_{11} &= P(A \cap B) \end{aligned}$$



The row and column probabilities are the eight possible conditional probabilities involving A and B , by the definition of conditional probability in (4):

$$\begin{aligned} n_{00}/n_{.0} &= P(A^*/B^*) = P(A^* \cap B^*) / P(B^*) \\ n_{10}/n_{.0} &= P(A/B^*) = P(A \cap B^*) / P(B^*) \end{aligned}$$

$$n_{01}/n_{.1} = P(A^*/B) = P(A^* \cap B) / P(B)$$

$$n_{11}/n_{.1} = P(A/B) = P(A \cap B) / P(B)$$

$$n_{00}/n_{0.} = P(B^*/A^*) = P(B^* \cap A^*) / P(A^*)$$

$$n_{01}/n_{0.} = P(B/A^*) = P(B \cap A^*) / P(A^*)$$

$$n_{10}/n_{1.} = P(B^*/A) = P(B^* \cap A) / P(A)$$

$$n_{11}/n_{1.} = P(B/A) = P(B \cap A) / P(A)$$

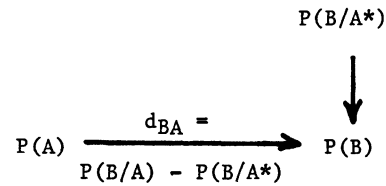
Now d_{BA} can be seen as one of the eight possible differences in row or column proportions, namely

$$d_{BA} = n_{11}/n_{1.} - n_{01}/n_{0.}. \quad (10)$$

This concept has a venerable tradition in contingency table analysis, particularly in the social sciences; hence the coefficient d_{BA} for the degree of $P(B)$ "explained" by event A will have intuitive appeal for analysts working in this tradition.

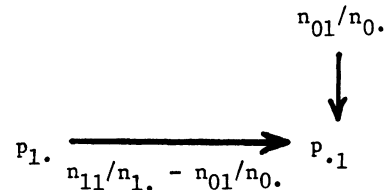
4. Causal Flow Graphs

Equations (8) and (9) can also be represented as a causal flow graph,



by applying four conventions: (1) probabilities of temporally prior or causative events (here $P(A)$) become values at source nodes, i.e., ones with outgoing arrows; (2) probabilities of dependent events or effects (here $P(B)$) become sink nodes, i.e., ones with incoming arrows; (3) the d_{ij} or "explained" probabilities (here d_{BA}) become coefficients of arrows running from source i to sink j ; and (4) "unexplained" probabilities (here $P(B/A^*)$) become "dummy" source nodes, i.e., ones with arrows running directly into a sink.

Causal flow graphs can be constructed directly from two-by-two tables using the following expressions:



Flow graphs have no unique mathematical properties; they merely translate equations like (8) into visual language. They do, however, facilitate substantive interpretations of data which might be less obvious in tabular or equation forms. The direct relationship between contingency tables, decomposition of conditional probabilities and causal flow diagrams has now been demonstrated.

5. Example with Two Events

The development of causal flow graphs owes much to the work of J.A. Davis (1975a). In order to facilitate comparisons between the conditional probability approach introduced here and the work of Davis, the data set used by him in illustrative examples will be adopted here. These data are pooled from the 1972, 1973 and 1974 National Opinion Research Center (NORC) General Social Surveys (GSS) of Americans age 18 and older; the sample sizes are 1613, 1504 and 1484, respectively, for a total of 4601.

To illustrate the two-event example discussed thus far,

A is living outside of the American South (in U.S. Census regions East and West South Central and South Atlantic) at age 16;

B is stated opposition to laws against marriages between Blacks and Whites.

The hypothesis is that upbringing outside of the South (A), an experience temporally prior to opinions expressed in 1972-4, constitutes a partial cause of tolerance on racial issues (B). (Black respondents, and those raised in foreign countries, or failing to answer one or both questions, are excluded from this example, thus lowering the sample size from 4601 to 3786).

The cross-tabulation of A and B, from Davis' published data, is

		B		
		0	1	
A	0	.550 597 (.158)	.450 489 (.129)	1086 (.287)
	1	.318 858 (.227)	.682 1842 (.486)	2700 (.713)
		.590 1455 (.384)	.790 2331 (.616)	3786 (1.000)

The four values required for the causal flow graph can be computed directly from this contingency table:

$$P(A) = p_{1.} = 2700/3786 = .713 \quad (11)$$

$$P(B) = p_{.1} = 2331/3786 = .616 \quad (12)$$

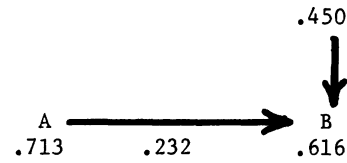
$$P(B/A^*) = n_{01}/n_{0.} = 489/1086 = .450 \quad (13)$$

$$d_{BA} = n_{11}/n_{1.} - n_{01}/n_{0.} = 1842/2700 - 489/1086 = .232 \quad (14)$$

Substituting in equation (8),

$$P(B) = .616 = \underset{\text{unexplained}}{.450} + \underset{\text{explained}}{(.232 * .713)} \quad (15)$$

The causal flow graph becomes:



It has been previously stated that such flow graphs facilitate substantive interpretations of data which might be less obvious in tabular or equation forms. The graph above, for example, might be given the following interpretation: Nonsouthern upbringing (A) is a partial cause of racial tolerance (B). A occurs with probability of .713 in the U.S.; when it does, the probability that B will also occur -- and would not have occurred otherwise -- is .232. B occurs with probability .616 -- with probability .450 in the absence of A, and with an additional probability of .166 (.713 * .232) as a result of A.

Worth noting here is the interpretation of d_{BA} in terms of conditional probability -- as an additional probability, or the probability that an A will produce a B that would not have otherwise occurred. This interpretation can be given formal statement:

d_{BA} is the probability that a B will accompany A that would not have occurred in the absence of A.

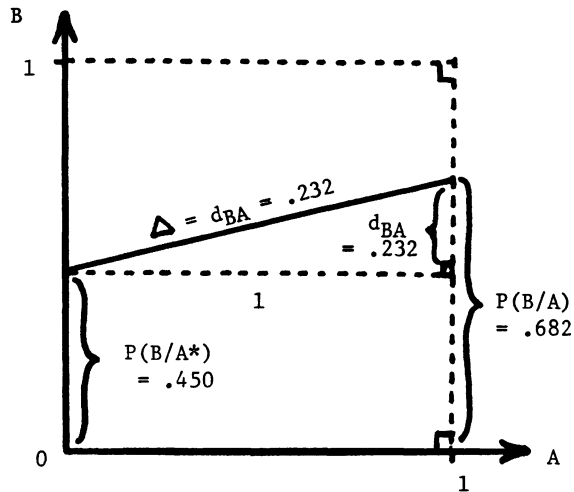
i.e., to repeat (9), $d_{BA} = P(B/A) - P(B/A^*)$. Because of its interpretation, d_{ij} is often termed the graph transmittance value from j to i.

6. Coordinate Plots

When binary variables like A and B are assigned the values 0 and 1, as in the tables here, at least three interpretations may be made in terms of coordinate plots: (1) the conditional probability of a 1-value on the dependent variable (B), given a 0-value on the independent variable (i.e., $P(B/A^*)$), is the y-intercept of a coordinate plot of B against A, or the constant in a linear equation like (8); (2) the conditional probability of a 1-value on B, given a 1-value on A (i.e., $P(B/A)$), is the intercept of line A = 1; and (3) the difference in proportions d_{BA} (i.e., $P(B/A) - P(B/A^*)$) is the slope of the linear relationship between B and A, or the coefficient in the linear equation. These graphic interpretations are illustrated in the coordinate plot at the head of the next page.

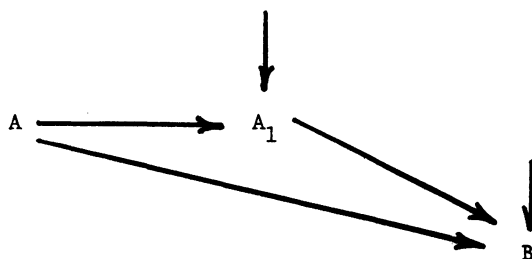
7. Three Events with No Interactions

The flow graph approach extends to systems with any number of events or variables. Consider the addition of a third event, A_1 , intervening in time between A and B. Again using the data set of Davis (1975a),



A_1 is educational attainment of at least a high school diploma.

The hypothesis is that, because upbringing outside the South (A) is more likely to result in high school education (A_1), this intervening event will at least partially "explain" the relationship between A and racial tolerance (B) in the Lazarsfeldian sense. The causal flow diagram of this hypothesis is



Because A is a direct partial cause of A_1 , the relationship between A and A_1 is the same as that between A and B in equation (8), namely,

$$P(A_1) = \underbrace{P(A_1/A^*)}_{\text{by A}} + \underbrace{d_{A_1A}P(A)}_{\text{by A}}, \quad (16)$$

where

$$d_{A_1} = P(A_1/A) - P(A_1/A^*). \quad (17)$$

Because B is partially caused by both A and A_1 , it is helpful--to keep the flow graph analysis relatively unencumbered in this example--to make a simplifying assumption, namely, that there are no interactions between A and A_1 in determining B. Stated formally, the assumptions of no interactions between A and A_1 are:

$$P(B/A \cap A_1) - P(B/A^* \cap A_1) = P(B/A \cap A_1^*) - P(B/A^* \cap A_1^*), \quad (18)$$

i.e., the direct effect of A on B is independent of A_1 , and

$$P(B/A \cap A_1) - P(B/A \cap A_1^*) = P(B/A^* \cap A_1) - P(B/A^* \cap A_1^*), \quad (19)$$

i.e., the direct effect of A_1 on B is independent of A. These assumptions are equivalent to the fact that d_{BA} is the same for both A_1 and A_1^* , and that d_{BA_1} is the same for both A and A^* .

Analogously to (9) and (17), because both A and A_1 are direct partial causes of B, and each has the same effect independent of the other factor (i.e., there are no interaction effects),

$$= P(B/A \cap A_1) - P(B/A^* \cap A_1) \quad (20)$$

$$d_{BA \cdot A_1} = P(B/A \cap A_1^*) - P(B/A^* \cap A_1^*),$$

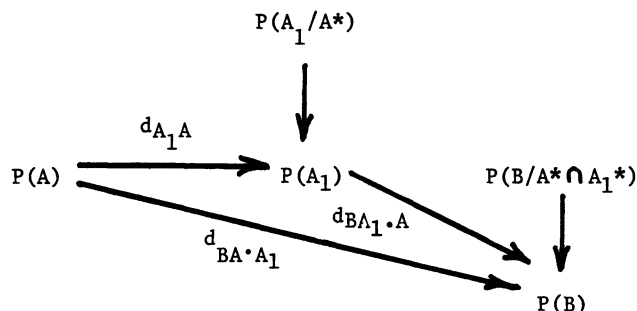
$$= P(B/A_1 \cap A) - P(B/A_1^* \cap A)$$

$$d_{BA_1 \cdot A} = P(B/A_1 \cap A^*) - P(B/A_1^* \cap A^*). \quad (21)$$

The complete equation for P(B), which combines (20) and (21), is

$$P(B) = \underbrace{P(B/A^* \cap A_1^*)}_{\text{by A or } A_1} + \underbrace{d_{BA \cdot A_1}P(A)}_{\text{by A}} + \underbrace{d_{BA_1 \cdot A}P(A_1)}_{\text{by } A_1}. \quad (22)$$

Equations (16) and (22) can be represented by the three-event causal flow graph



by applying the same conventions as above for source nodes and values, sink nodes, coefficients of arrows and "dummy" source nodes.

In terms of a three-way contingency table, marginal probabilities are the probabilities of individual events,

$$P_{0..} = P(A^*) \quad P_{.0.} = P(A_1^*) \quad P_{..0} = P(B^*)$$

$$P_{1..} = P(A) \quad P_{.1.} = P(A_1) \quad P_{..1} = P(B)$$

Cell probabilities are the probabilities of the eight possible intersections of A, A_1 and B,

$$P_{000} = P(A^* \cap A_1^* \cap B^*)$$

$$P_{100} = P(A \cap A_1^* \cap B^*)$$

$$P_{110} = P(A \cap A_1 \cap B^*)$$

etc.

$d_{BA \cdot A_1}$ and $d_{BA_1 \cdot A}$ depends on the simplifying assumptions (18) and (19), respectively, namely, that there are no interactions between A and A_1 in determining B. In other words, d_{BA} is independent of A_1 (or the same within categories of A_1),

$$d_{BA/A_1} = d_{BA/A_1^*}, \quad (33)$$

as shown in (27) and (28), and d_{BA_1} is independent of A_1 ,

$$d_{BA_1/A} = d_{BA_1/A^*}, \quad (34)$$

as shown in (29) and (30). When conditional ds are equal, as in (33) and (34), then d will have the same algebraic properties as coefficients in linear equations, or partial slopes in linear plots, as suggested by Section 6.

The interpretations of $d_{BA \cdot A_1}$ and $d_{BA_1 \cdot A}$ can be given formal statement:

$d_{BA \cdot A_1}$ is the probability that a B will accompany A that would not have occurred in the absence of A, independently of A_1 ;

$d_{BA_1 \cdot A}$ is the probability that a B will accompany A_1 that would not have occurred in the absence of A_1 , independently of A,

which is to repeat (20) and (21) in words. Because of this interpretation, $d_{ij.k}$ is often termed the graph transmittance value from j to i , controlling for (or "within categories of") k .

In analyzing actual data, assumptions (18) and (19) would be subject to empirical verification. There is good reason for the perfect fit (i.e., total lack of interaction) in Davis' data: he began with raw figures from the NORC surveys, estimated parameters and constants, tested for interactions (finding none to be significant), and then adjusted the data to fit the resulting model (1975a, p. 129). As he reports, "With small samples, data with no significant interactions can be bouncy; with large samples, models can fit quite well despite interactions that are statistically significant" (p. 130).

9. Discussion and Summary

Conditional probability serves to motivate an asymmetric interpretation of contingency tables based on differences in row and column proportions. This approach affords a natural causal interpretation, in the sense of changes in future probabilities, for linear flowgraph analysis of nominal or categorical variables like the "d system" of Davis (1975a). Motivation in terms of conditional probability will be particularly useful for survey and market researchers working with the elaboration model of the Lazarsfeldian school (involving "interpretation," reinforcers, suppressor variables, specification, spurious correlation, etc.; see Rosenberg 1968), and also as an introduction to path analysis for students familiar with statistical tables. Path diagrams involve variances and covariances, however, while flow graphs define absolute, partial and conditional probabilities. This difference makes

the flow graph approach more nearly like regression, particularly in the importance of asymmetric assumptions, and indeed (as suggested by Section 6) the two procedures give identical results for data free of interaction effects. For the treatment of such effects using flow graphs, the reader is referred to Davis (1975a, pp. 125-38).

10. References

- Bishop, Yvonne M.M., Stephen E. Fienberg and Paul W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.
- Coleman, James S. 1970. "Multivariate Analysis for Attribute Data." Pp. 217-45 in Sociological Methodology 1970, edited by E. Borgatta and G. Bohrnstedt. San Francisco: Jossey-Bass.
- Davis, James A. 1975a. "Analyzing Contingency Tables with Linear Flow Graphs: D Systems." Pp. 111-45 in Sociological Methodology 1976, edited by D. Heise. San Francisco: Jossey-Bass.
- _____. 1975b. "Communism, Conformity, Cohorts, and Categories: American Tolerance in 1954 and 1972-73." American Journal of Sociology 81: 491-513.
- Goodman, Leo A. 1963. "On Methods for Comparing Contingency Tables." The Journal of the Royal Statistical Society, Series A, 126: 94-108.
- _____. 1970. "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications." Journal of the American Statistical Association 65: 225-56.
- _____. 1972. "A General Model for the Analysis of Surveys." American Journal of Sociology 77: 1035-86.
- Heise, David R. 1975. Causal Analysis. New York: Wiley-Interscience.
- Huggins, William H., and Doris R. Entwisle. 1968. Introductory Systems and Design. Waltham, Mass.: Blaisdell.
- Kendall, Patricia L., and Paul F. Lazarsfeld. 1950. Pp. 133-96 in Continuities in Social Research: Studies in the Scope and Method of the American Soldier, edited by R.K. Merton and P.F. Lazarsfeld. New York: Free Press.
- Rosenberg, Morris. 1968. The Logic of Survey Analysis. New York: Basic Books.
- Stinchcombe, Arthur L. 1968. Constructing Social Theories. New York: Harcourt Brace Jovanovich.