1 Independence and association in the two-dimensional table

Introduction

This book describes models of category counts where each count is the observed frequency with which a unique combination of the levels of categorical variables occurs in a sample. This observed frequency provides an estimate of the probability of the variable-level combination in the population. The variable-level combinations are jointly displayed as a contingency table. Consider, for example, the simple two-way table formed by observing two categorical (as opposed to continuous) variables. Table 1.2 is a typical example with rows representing the categories of the variable residential area (A) and columns representing the categories of the variable preferred political party (B). For convenience, both A and B are treated as dichotomies, that is, each has two levels. More-complex data sets involving polytomies and more than two variables (and hence multiway tables) are investigated at length later; we remain with the two-way table in this first chapter to introduce some basic notation and concepts.

It is important at the outset to make explicit any assumptions that are implied by a methodology and, therefore, upon which the results of any analysis are conditional. The mere act of drawing up a contingency table represents a conscious structuring of a multifarious reality prior to analysis in which the variation present is subdivided according to distinct categorical variables. Although this may enable one to obtain plausible models, it is, it seems, an imposed world-view that may not be universally shared. Nonetheless, it is argued that, like the visual display unit, economic time series and scattergram, each of which is a device for data display, analogous to the contingency table, proffering its own salient information, the table of counts is an organising framework through which to view reality which is eminently suited to many practical situations.

Initially, raw data usually comprise a list, as in Table 1.1, which is only subsequently organised as a table. Of course, the advantage a contingency table like Table 1.2 has over a list is that it is a convenient and succinct summary of the data. It contains less information than the list from which it is derived since it does not tell us the order in which respondents were interviewed (assuming the list to be sequential), but in many studies this is redundant information and nothing is lost by ignoring it. Table 1.2 not only takes up less space than the complete list but also allows one to make a cursory examination of how frequently the various combinations of A and B occur, and to compare this set of numbers $\{f_{ij}\}$ with other sets $\{e_{ij}\}$ that would, on average, occur if various prespecified circumstances prevailed. In other words, the $\{e_{ij}\}$ are the expected frequencies generated according to some model. One of the most important models, or pre-suppositions about the mechanism generating the observed data, is that A and B are mutually independent. This will be discussed in some detail subsequently with respect to real data, and in later chapters we will consider other sets of expected data, and their allied models, which correspond to more-intricate propositions than simple independence.

It can be seen from Table 1.2 that the imaginary sample listed in Table 1.1 includes f_{21} rural Labour party 'affiliates' and that f_{ij} is the frequency in cell

Table 1.1. *Observations from which to construct a 2 × 2 contingency table*

Person number	Residential area (A)	Political party (B)
1	A_1	B_1
2	A_2	B_1
3	A_2	B_2
4	A_1	B_1
$\vdots$	$\vdots$	$\vdots$
f_{00}	A_2	B_2

key: f_{00} = total sample size.
A_1 = urban area.
A_2 = rural area.
B_1 = Labour party.
B_2 = not Labour.

Table 1.2. *Basic notation in a 2 × 2 contingency table*

	B_1	B_2	Total
A_1	f_{11}	f_{12}	f_{10}
A_2	f_{21}	f_{22}	f_{20}
Total	f_{01}	f_{02}	f_{00}

(i,j). Altogether there are f_{10} urban residents and f_{20} rural residents, whilst f_{01} are 'affiliated' to the Labour party and f_{02} are not.

Another presupposition prior to an analysis is that the variable levels employed in any table pertain to the true categories of a categorical variable and not some amalgamation of them. The effects of category amalgamation are discussed in a later chapter. The fact is that in many analyses the categories defining the table cells are only surrogates for true variable levels, and may be the result of coalescing some more finely divided categorisation or even a continuum. For example, residential areas may be perceived to assume positions on a graduated scale ranging from totally rural to totally urban and it is a considerable simplification to dichotomise this 'continuum', as in Table 1.2. However, this kind of simplification need not necessarily involve an information loss of the magnitude implied by the nominal-level representation of a continuous variable. There are important recent developments in contingency table analysis related to the logit model mentioned later (Goodman 1979*a*; Clogg 1982*b*; Duncan 1979) which take account of scores or quantitative values assigned to the levels of categorical variables. These are outlined in more detail in Chapter 5. Assigning a quantity to a category implies that the value is representative of all the individuals allotted to that category, though this may sometimes be too gross a generalisation with non-experimental data involving the uncontrolled variation of a continuous variable. It is true that in many cases the information loss incurred may be immaterial to the understanding of the system of interest, though, ultimately, the decision to opt for categorisation must be based on the costs and benefits involved, as perceived by the individual analyst.

A range of different models is available for data that cannot justifiably be reduced to tables of counts. Those models are not really separate from each other or from the models considered here, for links exist at the theoretical level to draw them together under the umbrella title of generalised linear models (Nelder & Wedderburn 1972; Nelder 1974; Dickinson 1977). It is generally acceptable to consider contingency table counts as Poisson (or, equivalently, multinomial (Birch 1963)) distributions, as described in Chapter 4. However, in the more general context, normal, gamma or other distributions may be appropriate. These comprise other types of generalised linear model, probably the most familiar of which is the classical linear regression model with normal errors.

A common algorithm exists for fitting many variants of generalised linear model, namely the method of iterative weighted least squares (see Chapter 5), and this provides the unifying framework with which to integrate otherwise disparate models. This algorithm forms the core of the interlinked GLIM and GENSTAT programs and gives immense scope in the selection of models that can be fitted (Baker & Nelder 1978). GENSTAT is used extensively in Chapters

5 and 6 for its capacity to easily provide the maximum likelihood estimates of the more specialised models encountered there. However, other forms of generalised linear model within the capacity of this package are not fitted. These comprise a vast field with an extensive literature which lies beyond the scope of this book.

While information loss may be a justifiable reason to reject the contingency table approach and opt for other types of generalised linear model, it is no longer true, as it was in the past, that contingency table analysis is an underdeveloped science from which only naive conclusions can be drawn, in contrast to the sophisticated multivariate analysis of continuous data. As a result of the considerable advances in the techniques of categorical data analysis, allied to the developing methodology outlined above (see Fienberg 1980), equally sophisticated analyses of either categorical or continuous variables, or both together, are now possible. The analyst of relations between categorical variables can now isolate the interactions between variables after allowing for other variables in the same way that the exponent of multiple regression can estimate partial regression coefficients. This is in stark contrast to the era when social, environmental and behavioural scientists were of necessity confined to the analysis of two-way tables, a restriction inviting fallacious inferences in the presence of multivariable associations.

Given that tables of counts are the focus of interest, there are a number of different approaches that could be adopted. This book does not attempt to cover all methods but a selection is made of what are judged to be of most practical value to the data analyst. Thus the emphasis is very much on the log-linear model, with some consideration also being given to the logit model. These interrelated approaches are by far the most versatile, comprehensive, and commonly used modelling procedures, though they in no way exhaust the available options. Hildebrand *et al.* (1977) and Bishop *et al.* (1975) describe some alternatives, and Fienberg (1980) traces various strands in the literature. Although estimation throughout is by maximum likelihood, alternatives exist which are described briefly in Chapter 5.

Independence in 2 × 2 tables

To establish more fully what is meant by this important proposition of independence, let us restrict ourselves initially to the simplest possible situation of two dichotomous variables such as those mentioned above. As will be seen, the discussion extends naturally to the $I \times J$ table.

It is illuminating to consider the quantity p_{ij}, the probability that an individual chosen at random from a population belongs to cell (i, j), even though in reality we would be presented with a set of numbers $\{f_{ij}\}$ and the set $\{p_{ij}\}$

could never be known exactly, given sample data. Nevertheless, the definition of independence is with respect to the $\{p_{ij}\}$ of Table 1.3 where strict equalities apply and where there is no need to consider the perturbing effects of random sampling variation.

Independence can be defined as the particular circumstance where knowing a person's A category is not going to help us guess his or her B category. Consequently,

$$p_{11}/p_{01} = p_{12}/p_{02} = p_{10}/p_{00} = p_{10}$$

and

$$p_{11} = p_{01}p_{10}$$

Likewise,

$$p_{11}/p_{10} = p_{21}/p_{20} = p_{01}$$

and

$$p_{11} = p_{10}p_{01}$$

If A and B are independent, then, generally, $p_{ij} = p_{i0}p_{0j}; i, j = 1, 2$. Furthermore, since $p_{11}/p_{12} =$ the odds on a B_1 response given that individuals are A_1, and $p_{21}/p_{22} =$ the odds on a B_1 response given that individuals are A_2, the independence condition requires that $p_{11}/p_{12} = p_{21}/p_{22}$.

In other words, independence means that the category proportions for B are unaltered by which A category is considered. Two equivalent ratios take a value of exactly 1.0 when there is independence. These are the odds ratio $(p_{11}/p_{12})/(p_{21}/p_{22}) = 1.0$, and the cross-product ratio, $p_{11}p_{22}/p_{21}p_{12} = 1.0$ (Mosteller 1968).

Table 1.3. *Theoretical probabilities in a 2 × 2 table*

	B_1	B_2	Total
A_1	p_{11}	p_{12}	p_{10}
A_2	p_{21}	p_{22}	p_{20}
Total	p_{01}	p_{02}	p_{00}

$$\sum_i p_{ij} = p_{0j}$$

$$\sum_j p_{ij} = p_{i0}$$

$$\sum_i \sum_j p_{ij} = p_{00} = 1.0$$

It has already been stated that the $\{p_{ij}\}$ are theoretical probabilities and are unknown, though they can be estimated from sample data since $\hat{p}_{ij} = f_{ij}/f_{00}$, where $\hat{p}_{ij}$ signifies the maximum likelihood estimate of p_{ij}.

If A and B are independent, then we know that $p_{ij} = p_i p_j$; since $\hat{p}_{i0} = f_{i0}/f_{00}$ and $\hat{p}_{0j} = f_{0j}/f_{00}$, $\hat{p}_{ij} = f_{i0}f_{0j}/f_{00}^2$ and hence $e_{ij} = f_{00}\hat{p}_{ij} = f_{i0}f_{0j}/f_{00}$.

Given a total of f_{00} observations in the table, the expected frequency in the (i, j) cell is e_{ij}, given independence. A comparison of the set of expected frequencies $\{e_{ij}\}$ with the set observed $\{f_{ij}\}$ provides the basis of a test of the independence assumption. Table 1.4 contains these frequencies.

We note that the set of observed frequencies in Table 1.4 differs somewhat from that to be expected were the variables independent, though we cannot be sure at this point whether the difference between $\{f_{ij}\}$ and $\{e_{ij}\}$ should be assigned to sampling variation or whether it represents a consistent deviation that would be repeated in other samples. In order to choose between these optional interpretations we utilise the familiar chi-squared test of independence. In fact, we adopt two asymptotically equivalent formulae for our test statistic, since the less familiar of the two allows us to discuss in simple terms one very important issue that is relevant to all the tests performed throughout this book.

The two formulae are:

$$X^2 = \sum_i \sum_j (f_{ij} - e_{ij})^2 / e_{ij}$$

$$X^2 = [\log(f_{11}f_{22}/f_{12}f_{21})]^2 / (f_{11}^{-1} + f_{12}^{-1} + f_{21}^{-1} + f_{22}^{-1})$$

Using the first formula, $X^2 = [(21 - 8.76)^2/8.76] + [(25 - 37.24)^2/37.24] + [(3 - 15.24)^2/15.24] + [(77 - 64.76)^2/64.76] = 33.3$, a highly unlikely observation from the χ^2 distribution to which X^2 would approximate were noise perception and pressure-group support independent. Conventionally, we would accept

Table 1.4. *Observed frequencies and corresponding expected frequencies assuming independence*

	Observed			Expected		
	B_1	B_2	Total	B_1	B_2	Total
A_1	21	25	46	8.76	37.24	46
A_2	3	77	80	15.24	64.76	80
Total	24	102	126	24	102	126

key: A_1 = respondent supports anti-Stansted airport pressure group.
A_2 = respondent does not support pressure group.
B_1 = respondent considers the local area noisy.
B_2 = respondent does not consider the local area noisy.

any value of $X^2 \geqslant \chi^2_{1,\,0.05} = 3.84$ as being sufficiently unlikely, thus suggesting that the variables are related, though of course such high values could occasionally occur by chance (with probability $\alpha = 0.05$) since 'freak' samples can occur spuriously suggesting association when none in fact exists in the population as a whole. Such a mistaken inference is referred to as a type I error. We could reduce this probability α (the probability of a type I error or level of risk) to, say, 0.01, in which case the critical value would then be $\chi^2_{1,\,0.01} = 6.64$, though this would be at the cost of increasing the probability of a type II error, which occurs when the sample deceptively suggests independence.

The alternative and less familiar formula for X^2 will usually produce similar values to those generated by the above formula in large samples since both statistics are distributed asymptotically as χ^2. We now find that

$$X^2 = \{\log\,[(21 \times 77)/(25 \times 3)]\}^2/$$
$$[(1/21) + (1/25) + (1/3) + (1/77)] = 9.43/0.43 = 21.73,$$

which, although somewhat less than 33.3, also has a very small probability of occurring in the χ^2_1 distribution.

Note that the numerator of this equation is a function of the cross-product ratio $(f_{11}f_{22}/f_{12}f_{21})$ defined above with probabilities, where we established that it will assume a value 1.0 with independent variables. Thus log $(f_{11}f_{22}/f_{12}f_{21})$ has a mean of 0, assuming independence, and since it has an approximately normal distribution it will lie within 1.96 standard deviations of this mean in 95% of random samples. The standard deviation is the square root of the denominator in the above X^2 equation. In fact, the relation between the normal and chi-squared distributions is illuminated by this equation since it shows that χ^2 is simply the square of a normally distributed variable with mean 0 and variance 1.0.

The denominator is a function of the observed cell frequencies and it will increase in value as they diminish. What this in effect means is that, given two samples of different size, ostensibly displaying the same degree of association as measured by the sample cross-product ratio $f_{11}f_{22}/f_{12}f_{21}$, one could very easily conclude that there is evidence of association in the population on the basis of the larger sample and its resulting X^2 value, whereas perhaps X^2 would not be deemed an unusual observation from a χ^2_1 distribution in the smaller sample due to the inflation of the denominator, which consequently deflates X^2.

This is illustrated in a concrete way by Table 1.5 which describes a smaller sample drawn from the same Stansted population as the data in Table 1.4. Both tables display an equivalent degree of association between pressure-group support and noise perception, on the basis of direct comparison of respective cross-product ratios, but Table 1.5 provides less-convincing (though still highly significant) statistical evidence of non-independence.

From Table 1.5 we find that $X^2 = \{\log [(5 \times 20)/(2 \times 2)]\}^2/[(1/5) + (1/2) + (1/2) + (1/20)] = 10.36/1.25 = 8.29$ and that $P(\chi_1^2 > 8.29) = 0.004$, as opposed to $P(\chi_1^2 > 21.73) = 0.000\,003$, for the larger sample in Table 1.4. The small observed frequencies of Table 1.5 raise questions about the validity of the assumption that X^2 approximates to χ^2, given independence, since they produce one distinctly small expected value. The question of the validity of the X^2 approximation to χ^2 for the 2 x 2 table has been considered by a number of authors, commencing with Yates (1934) who introduced a continuity correction giving $X^2 = \Sigma_i \Sigma_j (|f_{ij} - e_{ij}| - \frac{1}{2})^2/e_{ij}$. This is an often-used formula, though Fienberg (1980) reports that it may increase the probability of a type II error when referred to the χ_2^1 distribution. We refrain from using it here. Fisher's exact test (see Upton 1978) is an alternative approach, though difficult to generalise to the multiway table. Small samples are discussed in the context of the multiway table in Chapter 5. The converse problem, inference from large samples, is considered in Chapter 6.

The main point of the current discussion is the possibility that our conclusions about the presence or absence of association depend on the size of the sample. Although the chosen measure, the (log) cross-product ratio, will still, on average, take a value one (zero) in the absence of association whatever the sample size, the normal range of its variation around this expected value will not be stable. This is illustrated by comparing the square roots of the denominators used to calculate X^2 for Tables 1.4 and 1.5, $\surd(0.43)$ and $\surd(1.25)$ respectively. Thus, although the cross-product ratio is marginally larger for Table 1.5, it could more easily have been generated by sampling from a population in which the variables are independent.

The above illustrates that a large sample provides more evidence, or information, on which to base conclusions about a population, and this generalisation is true for the variety of hypotheses tested throughout this book. This is why, formally, we can never accept a null hypothesis, which, in the present context, is the statement that there is no association between the variables. The most we can do is fail to reject it on the grounds of insufficient evidence of the existence

Table 1.5. *A small sample from the Stansted population*

	B_1	B_2
A_1	5	2
A_2	2	20

key: as in Table 1.4.

of association in the population from which the sample is drawn. One can never prove conclusively that no association exists in the population simply because another analyst may take a larger, more revealing, sample which provides the evidence we failed to detect.

The above comments are made with reference to the simple two-way table and data generated by the independence model, though they are equally applicable to the larger tables and more-complex models to be encountered in later chapters. A further comment on these and related issues is to be found later in the book after some of the details of these complexities have been made more explicit.

The $I \times J$ table

The notation and concepts applicable to the 2 × 2 table are easily extended to the $I \times J$ table which summarises data such as that in Table 1.6. The essential difference between Table 1.6 and Table 1.1 is that the former describes the variation of two polytomous variables rather than two dichotomies. The essence of this variation would be discernible from a two-way contingency table such as Table 1.7.

If A and B were independent, we would find that $f_{ij}/f_{i0} \approx f_{0j}/f_{00}$, where f_{ij} is the observed frequency in typical cell (i,j). Although exact equalities are unlikely because of random sampling variation, this approximate equality would hold for all i, j. In terms of theoretical cell probabilities, independence can again be defined as $p_{ij} = p_{i0}p_{0j}$, $i = 1, 2, \ldots, I; j = 1, 2, \ldots, J$.

Table 1.6. *Observations from which to construct an $I \times J$ contingency table*

Person number	Residential area (A)	Political affiliation (B)
1	A_3	B_6
2	A_1	B_2
3	A_1	B_2
⋮	⋮	⋮
f_{00}	A_2	B_4

key: A_1 = inner city.
A_2 = suburbs.
A_3 = rural.

B_1 = Labour party.
B_2 = Liberal party.
B_3 = Social Democrats.
B_4 = Conservative party.
B_5 = other parties.
B_6 = unaffiliated.

Table 1.8 is an $I \times J$ contingency involving two polytomous variables, car availability and fatigue, compiled from data originally analysed by Bowlby & Silk (1982). We find that $X^2 = \Sigma_i \Sigma_j (f_{ij} - e_{ij})^2/e_{ij} = 23.42 > \chi^2_{8,0.05} = 15.51$, indicating significant association between car availability and fatigue. This is a rather general statement in which no attempt is made to identify the source of the interaction between the two variables. These data are discussed from another viewpoint in relation to the saturated log-linear model later in this chapter, and in Chapter 5 we attempt to make more-precise statements about the way in which pairs of polytomous variables interact. Note that the enlarged critical value $\chi^2_{8,0.05}$ reflects the fact that the independence model is here fitted

Table 1.7. *Basic notation in an* $I \times J$ *contingency table*

	B_1		B_2	B_3		B_4	B_5	...	B_J	Total
A_1	f_{11}		f_{12}	f_{13}	...				f_{1J}	f_{10}
A_2	f_{21}		f_{22}	f_{23}	...				f_{2J}	f_{20}
A_3	f_{31}		f_{32}	f_{33}	...				f_{3J}	f_{30}
⋮									⋮	⋮
A_I	f_{I1}	...							f_{IJ}	f_{I0}
Total	f_{01}		f_{02}	f_{03}	...				f_{0J}	f_{00}

Table 1.8. *An* $I \times J$ *contingency table from a shopping survey in Oxford*

	B_1	B_2	B_3	B_4	B_5	Total
A_1	55	11	16	17	100	199
A_2	101	7	18	23	103	252
A_3	91	20	25	16	77	229
Total	247	38	59	56	280	680

key: A_1 = no car availability.
A_2 = some car availability.
A_3 = full car availability.

The level assigned to respondents on the fatigue variable B is determined by the extent of their agreement with the assertion: 'I find getting to grocery shops very tiring.'

B_1 = disagree.
B_2 = tend to disagree.
B_3 = in between.
B_4 = tend to agree.
B_5 = agree.