THE BAYESIAN ANALYSIS OF CONTINGENCY TABLES

By Dennis V. Lindley

University College of Wales and Harvard Business School

Summary. This paper describes how data from a multinomial distribution, and in particular data in the form of a contingency table, may be studied by using a prior distribution of the parameters and expressing the results in the form of a posterior distribution, or some aspects thereof, of the parameters. The analysis used must depend on the prior distribution and the form described here only applies to a certain type of prior knowledge but, for reasons given below, it is believed that this type is of frequent occurrence. The binomial situation is first considered and the results obtained there suggest a general result for the multinomial distribution, which is then established. A few remarks on Bayesian analysis in general enable the result to be applied, first to certain multinomial problems and then, with the aid of another general result, to contingency tables. The method used there has close connections with the Analysis of Variance and these connections are examined, particularly with a view to simplifying the analysis of contingency tables involving three or more factors.

1. Binomial distributions. Although it will appear as a special case of results to be established for the general multinomial situation, it is instructive to begin with the binomial distribution which suggested the generalizations. Let N independent trials with constant probability θ of success result in n successes and $(N - n)$ failures. The likelihood is

$$\theta^n(1 - \theta)^{N-n}. \tag{1.1}$$

As other authors have remarked, it is convenient to take as prior distribution one with density proportional to

$$\theta^a(1 - \theta)^b \tag{1.2}$$

for suitable $a, b > -1$. The limiting case where a and b both tend to -1 is also important. The posterior density is then proportional to

$$\theta^{a+n}(1 - \theta)^{b+N-n}. \tag{1.3}$$

The known relation between this beta-distribution and the F-distribution enables the last result to be expressed by saying that

$$F = \left(\frac{b + N - n + 1}{a + n + 1}\right)\left(\frac{\theta}{1 - \theta}\right) \tag{1.4}$$

has an F-distribution on $[2(a + n + 1), 2(b + N - n + 1)]$ degrees of freedom. When Fisher first introduced the distribution he suggested using the z-transfor-

Received 30 December 1963; revised 4 August 1964.

mation of it on the grounds that z was more nearly normally distributed than was F.

Since the distribution of the ratio of two independent χ^2-variables is F, the use of the logarithmic transformation of F and its approximate normality is intimately related to the same transformation of a χ^2, or gamma, variable and its approximate normality. The latter transformation has been discussed by Bartlett and Kendall [2] who remark that the "transformation may safely be used for $n = 10$ and over, more tentatively from $n = 5$ to $n = 9$, and probably not at all below $n = 5$." (Here n is the degrees of freedom for χ^2.) In our context the degrees of freedom are not less than twice the number of successes (or of failures) so that the transformation can certainly be used provided the smaller of these two numbers is 5 or more and probably when 3 or 4. (The recommendation is similar to that for the χ^2-approximation to Pearson's goodness-of-fit statistic which, as we shall see below, we recommend be replaced by another statistic. We have more to say about this at the end of Section 3.)

Fisher used $z = \frac{1}{2} \ln F$: in this context it is more convenient to omit the half and use the natural logarithm of F. Fisher's approximation then amounts to saying that if F has ν_1 and ν_2 degrees of freedom then $\ln F$ is approximately normal with mean $\ln \{(\nu_1 - 1)/(\nu_2 - 1)\}$ and variance $2(\nu_1^{-1} + \nu_2^{-1})$. In both moments the terms omitted are of order ν_1^{-2}, $(\nu_1\nu_2)^{-1}$ and ν_2^{-2}. For $\ln\{\theta/(1 - \theta)\}$ the mean is therefore

$$\ln \{(a + n + \tfrac{1}{2})/(N - n + b + \tfrac{1}{2})\}, \tag{1.5}$$

and the variance

$$(a + n + 1)^{-1} + (b + N - n + 1)^{-1}; \tag{1.6}$$

the approximations being good for large $a + n$ and $b + N - n$, probably above 5.

One is thus led to consider the natural logarithm of the odds, $\theta/(1 - \theta)$, in favour of success. The initial reason for using it here is its convenient approximate posterior distribution, but reflection suggests that it is a convenient quantity to consider for other reasons. If the value of θ were known, and less than $\frac{1}{2}$, it would be usual to quote its value to a fixed number of decimal places: the argument being that it is proportional changes in θ that are typically of interest. In these circumstances the logarithm of θ could be quoted to a fixed number of places and a given change in it would be of equal significance for all θ. But the argument should be symmetric in success and failure and hence the logarithm of the odds, or simply the log-odds, can be used. A further, and more important argument in favour of using the log-odds is the existence of certain additive properties that they have in situations to be discussed below: essentially these arise from the fact that the important property of independence is expressed multiplicatively and therefore, in terms of logarithms, additively. The results of this paper will be expressed entirely in terms of log-odds and their multinomial generalizations.

We next consider the prior distribution (1.2). We propose to consider the limiting case $a = b = -1$ and justify this choice by two main arguments as follows. In many situations the prior knowledge of θ is small, and there is some advantage in considering an analysis for a standard prior which could be used in most applications where no appreciable prior knowledge is available. If the log-odds is to be used then the original Bayes hypothesis would suggest taking all values of it to be equally likely. This gives $a = b = -1$. That this is sensible can be seen by noting that the results of trials cannot decrease the powers of θ and $(1 - \theta)$ in the beta-distribution and presumably cannot decrease the amount of information about θ. Consequently as small values as possible for a and b would correspond to the least prior information about θ. In order that the prior distribution be proper (i.e. integrate to 1) it is necessary that a and b both exceed -1 so that $a = b = -1$ is a lower bound, though not attained, within the class of proper prior distributions. The resulting distribution is improper but this need not concern us since we will be using approximations where the degrees of freedom for F are large and the posterior distributions are proper.

A second reason for restricting attention to $a = b = -1$ is that this special case can be regarded as a canonical form for all prior distributions belonging to the beta-family (1.2). By this we mean that any beta-distribution can be reduced to the canonical form $a = b = -1$. To see this we remark that if the prior distribution is given by (1.2) it may be regarded as a posterior distribution to data consisting of $(a + 1)$ successes and $(b + 1)$ failures, this data having prior distribution with $a = b = -1$. Consequently if the actual data yield n successes and $(N - n)$ failures we may regard the total knowledge of θ to consist of $(n + a + 1)$ successes and $(N - n + b + 1)$ failures for which the prior has the canonical form. Consequently by adding $(a + 1)$ and $(b + 1)$ respectively to the numbers of successes and failures actually observed the beta-distribution may be replaced by the canonical form. Of course, the method is not available for prior distributions which are not beta: see remarks in Section 8 below.

In many applications it will probably be true that the values of a and b for the prior distribution will be small. If the sample values, n and $(N - n)$, the numbers of successes and failures, are large, we see from (1.5) and (1.6) that the actual small values of a and b will not be important. It is therefore reasonable to take $a = b = -1$. With $a = b = -1$ the posterior distribution of $\ln \{\theta/(1 - \theta)\}$ is approximately normal with mean

$$\ln \{(n - \tfrac{1}{2})/(N - n - \tfrac{1}{2})\} \tag{1.7}$$

and variance $n^{-1} + (N - n)^{-1}$. For notational simplicity the $\frac{1}{2}$'s which occur in (1.7) will be omitted, with little loss in accuracy, and we shall write

$$\ln \{\theta/(1 - \theta)\} \sim N[\ln \{n/(N - n)\}, n^{-1} + (N - n)^{-1}]. \tag{1.8}$$

Throughout the paper some improvement in the approximations for the prior with $a = b = -1$ may be effected by reducing the observations by $\frac{1}{2}$ in calculating the posterior means. In words (1.8) says that the true log-odds is approximately normally distributed about the sample log-odds with variance equal to the sum of the reciprocals of the numbers in the two classes, success and failure.

The remainder of the paper is concerned with generalizations of the result in the last sentence, in particular with the meaning of log-odds in the multinomial situation, and with applications.

2. Multinomial distributions. In the multinomial situation a slightly different notation is desirable. Let k denote the number of classes: in the binomial case $k = 2$. Let $\theta_1, \theta_2, \cdots, \theta_k$ denote the probabilities for the k classes: necessarily $\theta_i \geqq 0$ for all i, and $\sum \theta_i = 1$ (all summations in this section run from 1 to k). Let $n_1, n_2, \cdots, n_k$ denote the observed numbers in each of the classes when $N = \sum n_i$ independent trials are made with the above probabilities for the classes.

We use another suggestion of Fisher's. He remarked that if $n_i (i = 1, 2, \cdots k)$ were independent Poisson variables with means Ψ_i, then the conditional distribution of them, given $N = \sum n_i$, would be multinomial as just described with $\theta_i = \Psi_i / \sum \Psi_i$. The proof is immediate since N has a Poisson distribution with mean $\sum \Psi_i$ and hence

$$\begin{aligned} p(n_1, n_2, \cdots, n_k \mid N) &= p(n_1, n_2, \cdots, n_k)/p(N) \\ &= \frac{e^{-\Sigma\Psi_i} \prod(\Psi_i^{n_i}/n_i\,!)}{e^{-\Sigma\Psi_i}(\sum\Psi_i)^N/N!} \qquad (2.1) \\ &= N!\prod(\theta_i^{n_i}/n_i\,!)\,, \end{aligned}$$

as required. An alternative way of regarding this result is to remark that the probability distribution of the Poisson n_i factors into the distribution of N, which depends only on $\theta = \sum \Psi_i$, and the conditional multinomial distribution of the n_i given N, which depends only on $\theta_1, \theta_2, \cdots, \theta_k$. If the prior distribution of the Ψ_i similarly factors into one part which depends only on θ and another which depends only on the θ_i, the same will be true of the posterior distribution. Consequently, under these conditions on the prior, the posterior distribution of the θ_i will depend only on the multinomial part of the likelihood. Thus this posterior may be obtained by the Poisson device.

An appropriate prior distribution for the Ψ_i may be obtained by supposing that their logarithms are independent and uniformly distributed over the whole real line. This prior distribution factors in the way described above. The Jacobian of the transformation from $\Psi_1, \Psi_2, \cdots, \Psi_k$ to $\theta, \theta_1, \theta_2, \cdots, \theta_{k-1}$ is easily found to be θ^{k-1}, whence

$$\frac{d\Psi_1 d\Psi_2 \cdots d\Psi_k}{\Psi_1\Psi_2 \cdots \Psi_k} = \theta^{k-1}\frac{d\theta d\theta_1 \cdots d\theta_{k-1}}{\theta^k\theta_1 \cdots \theta_k} = \left(\frac{d\theta}{\theta}\right)\left(\frac{d\theta_1 \cdots d\theta_{k-1}}{\theta_1 \cdots \theta_k}\right) \qquad (2.2)$$

as required. The posterior distribution of the θ_i is then obtained by multiplying the last factor in (2.2) by the likelihood from (2.1) with the result proportional to

$$\prod \theta_i^{n_i - 1}. \qquad (2.3)$$

We seek an approximation to (2.3). We find this by considering the Poisson distribution. The posterior distribution of the Ψ_i, given the n_i (now Poisson variables) is proportional to

$$\prod (e^{-\Psi_i}\Psi_i^{n_i-1}). \tag{2.4}$$

Thus the Ψ_i are independent and each is distributed in a Type III or Gamma distribution. But, as explained above, if a variable has such a distribution, then its logarithm is approximately normally distributed for all but small values of n_i. It is easy to calculate the mean and variance with the result that the $\ln \Psi_i$ are independent and approximately normally distributed with means $\ln n_i$ and variances n_i^{-1}.

Let $a_1, a_2, \cdots a_k$ be a set of constants with $\sum a_i = 0$. In common with the nomenclature used in the design and analysis of experiments, a linear form in quantities with coefficients which add to zero will be termed a *contrast* in those quantities. Consider a contrast in the $\ln \Psi_i$.

$$\sum a_i \ln \Psi_i = \sum a_i \ln (\theta\theta_i) = \sum a_i \ln \theta_i$$

since $\sum a_i = 0$. Consequently a contrast in the $\ln \Psi_i$ is equally a contrast in the $\ln \theta_i$. But the contrasts in the $\ln \Psi_i$ are approximately normally distributed and any set of them is approximately jointly normally distributed. The same must therefore apply to the $\ln \theta_i$ and, by the arguments given earlier, these distributions of the $\ln \theta_i$ apply to multinomial sampling. Consequently we have the following

THEOREM 1. *If the random variables $n_1, n_2, \cdots, n_k$ have a multinomial distribution with parameters $\theta_1, \theta_2, \cdots, \theta_k$; and if the prior distribution of the θ_i has density proportional to $(\prod \theta_i)^{-1}$ over the region $\theta_i \geqq 0$, $\sum \theta_i = 1$: then if the constants a_{pi} $(p = 1, 2, \cdots, m; i = 1, 2, \cdots, k; m < k)$ satisfy $\sum_i a_{pi} = 0$, the joint posterior distribution of the contrasts $\sum_i a_{pi} \ln \theta_i$ $(p = 1, 2, \cdots, m)$ is approximately normal with means*

$$\sum_i a_{pi} \ln n_i \tag{2.5}$$

and covariances (variances when $p = q$)

$$\sum_i a_{pi}a_{qi}n_i^{-1}. \tag{2.6}$$

The expressions for the means and covariances follow from the independence of the $\ln \Psi_i$ and their means and variances.

The means given by (2.5) may be written

$$\sum_i a_{pi} \ln (n_i/N), \tag{2.7}$$

since $\sum_i a_{pi} = 0$. This may be preferred since the individual terms in the sum are then of the same order as those in the corresponding contrast $\sum_i a_{pi} \ln \theta_i$: θ_i being of the same order as n_i/N.

In the binomial case, $k = 2$, the only contrasts are multiples of $\ln \theta_1 - \ln \theta_2 = \ln \{\theta/(1 - \theta)\}$, the log-odds, in the notation of Section 1. Consequently the earlier result for the binomial is a special case of the theorem. In view of the generalization we shall call a contrast in the $\ln \theta_i$ a *log-odds*, despite the fact that the odds do not necessarily refer to a single contrast of one event with another. Thus $\ln \theta_1 - \ln \theta_3$, for example, is the logarithm of genuine conditional odds, θ_1/θ_3, but we shall use the name for $\ln \theta_1 - 2 \ln \theta_2 + \ln \theta_3$ which has not this property. The theorem can now be expressed as saying that the approximate posterior distribution of the log-odds is as it would be if the $\ln \theta_i$ were approximately $N(\ln n_i, n_i^{-1})$ and independent. (Of course, the $\ln \theta_i$ are not themselves even approximately independent since $\sum \theta_i = 1$.) Consequently if attention is confined to log-odds then the simple normal result may be used: in particular techniques of the Analysis of Variance are available. We shall see below that many parameters of interest in the analysis of multinomial data, particularly in the form of contingency tables, are expressible as log-odds, and that therefore the normal theory can be applied. As in the binomial case the approximation can be improved by subtracting $\frac{1}{2}$ from n_i in (2.5).

3. Remarks on Bayesian analysis. Before proceeding to discuss application of Theorem 1, it is necessary to clarify a few points in statistical analyses that overtly use Bayes theorem and a prior distribution. The object of such an analysis is to provide the posterior distribution of the parameters or, if only some of them are of interest, the marginal posterior distribution of those. If the data are subsequently to be used as a basis for decision then the posterior distribution provides the necessary material for the calculation of the best decision. Once the posterior distribution has been obtained the only problem remaining is the descriptive one of how to present it. If we follow classical statistics and its concept of a confidence interval, it is rather natural to summarize the posterior distribution by giving an interval which contains a rather large, say 95%, amount of the posterior probability. This has the disadvantage of only giving information about the tails of the distribution and does not provide, for example, any good idea of the most probable values. The provision of the median and other quantiles in addition to the 95% values would help. For some purposes even an interval may be thought to be an overelaborate summary, particularly in cases where one value of the parameter is of especial interest. In these cases it may be felt adequate to say whether this value is a probable one on the basis of the posterior distribution. Again borrowing an idea from classical statistics, where the confidence set is the set of values that would not be rejected if tested as null hypotheses in a significance test at the level of the confidence set, we can perform a Bayesian significance test by seeing whether the particular value lies within the interval containing proportion $(1 - \alpha)$ of the posterior distribution. If not, then it can be said that this value is significant at level α: in the sense that it does not belong to a set of values having reasonably high posterior probability.

Of course there are many intervals containing a given proportion of the posterior distribution. We shall choose that interval (typically unique) which is such that no values within it are less probable than any value without it. This provides the shortest interval in the usual sense. These ideas are discussed in much more detail in [15] but we would point out here that no account has been taken, in formulating this concept of a significance test, in any economic or general decision-theoretic considerations. This seems to be in the same spirit as a classical significance test. Neither have we incorporated strong prior ideas about the value of especial interest. Our prior distributions are "smooth" in the neighbourhood of the special value. This is in marked contrast to Jeffreys' significance tests [10] where there is a concentration of prior probability on the special value.

We now apply these ideas to the posterior distributions of log-odds. The normal posterior distribution is particularly easy to understand for a single parameter. For several parameters the multivariate form is not so simple to comprehend. Let $\phi_1, \phi_2, \cdots, \phi_s$ be s linearly independent log-odds with means m_i and covariances v_{ij}, say. The expressions for these means and covariances follow from (2.5) and (2.6). The joint density of the ϕ_i is constant on the ellipsoids

$$\sum_{i,j=1}^{s} (\phi_i - m_i)v^{ij}(\phi_j - m_j) = c, \tag{3.1}$$

where v^{ij} are the elements of the inverse of the dispersion matrix, and c is any positive constant Furthermore the left-hand side of (3.1) is distributed as χ^2 on s degrees of freedom. If χ^2_α is the upper $100\alpha\%$ point of this distribution, the posterior probability that

$$\sum_{i,j=1}^{s} (\phi_i - m_i)v^{ij}(\phi_j - m_j) \leqq \chi^2_\alpha \tag{3.2}$$

is $(1 - \alpha)$. With $\alpha = 0.05$, the posterior probability of (3.2) is 95% but, of course, any value of α may be used. In particular if the values $\phi_i = \phi_i^{(0)}$ are of interest, they would seem unlikely if they did not satisfy (3.2) with $\phi_i = \phi_i^{(0)}$ $(i = 1, 2, \cdots, s)$. This provides a significance test of the hypothesis that $\phi_i = \phi_i^{(0)}$. If $\phi_i^{(0)} = 0$ $(i = 1, 2, \cdots, s)$ the relevant statistic reduces to

$$\sum_{i,j=1}^{s} m_i v^{ij} m_j \tag{3.3}$$

and may be compared with χ^2.

In applications of (3.2) and (3.3) it should be remembered that although the m_i and v^{ij} are statistics and the ϕ_i are parameters, it is the ϕ_i that are the random variables. In the usual argument, wherein the m_i are random variables, normally distributed about means ϕ_i with covariances v_{ij} (supposed known), statements like (3.2) and (3.3) are still correct. In particular the comparison of (3.3) with the χ^2 distribution is basic to the Analysis of Variance, the variance being supposed known. Consequently in analyzing the log-odds by the Bayesian methods just described, we have available the methods of the Analysis of Variance. In the following sections we will continually make use of ideas derived from that classical field.

Before proceeding to the applications of Theorem 1, two remarks need to be made in the light of these comments on Bayesian analyses. The first concerns the invariance of methods based on (3.2) and (3.3) under linear transformations of the log-odds. This is because the quadratic forms are so invariant. A consequence of this remark is the observation that it is possible to replace any particular set of log-odds by linear transformations thereof which may simplify the analysis.

The second remark concerns the approximation to the Type III distribution (Equation (2.4)) that is basic to the methods. There are, of course other, and better, approximations: for example, the Wilson-Hilferty one using $\chi^{\frac{2}{3}}$; but they cannot compare in convenience with the log-odds. One approximation that does merit serious attention is the classical χ^2-approximation based on the statistic $\sum(O - E)^2/E$ in an obvious notation. This can be used in a Bayesian analysis along the same lines as just explained for (3.3): details are given in [15]. It is hoped to compare the results obtained from the χ^2 and log-odds approaches in another paper. Preliminary numerical work suggests that χ^2 is worse than the log-odds in supplying an approximation to the posterior distribution. This is not surprising in view of the fact that χ^2 was not designed to do this. However, if a uniform prior distribution for θ in the binomial case replaces the uniform for $\ln \theta$ used above, the effect being to increase the n_i by 1 in the log-odds results, then χ^2 seems to be a good approximation to the posterior distribution.

4. Certain multinomial problems.

In the multinomial situation described in Section 2, let $\theta_i = \theta_i^{(0)}$ $(i = 1, 2, \cdots k)$ be a simple null hypothesis specifying the values of all the parameters. We now proceed to derive a test of this null hypothesis using the above results on the posterior distribution of log-odds and the principle of Bayesian analysis just outlined. The posterior density of the $\ln \Psi_i$ (from the Poisson distribution) is proportional to

$$\exp\{-\tfrac{1}{2}\textstyle\sum(\ln \Psi_i - \ln n_i)^2 n_i\}, \tag{4.1}$$

using the approximation. This distribution will also apply to the $\ln \theta_i$ provided that only contrasts are considered. Now (4.1) may be rewritten, using the usual breakdown of a sum of squares into a sum of squares about a weighted mean plus a term involving the mean. Let $u_i = \ln \Psi_i - \ln n_i$, normally distributed about zero with variance n_i^{-1}; then (4.1) is

$$\exp\{-\tfrac{1}{2}[\textstyle\sum n_i(u_i - \bar{u})^2 + N\bar{u}^2]\} \tag{4.2}$$

with $\bar{u} = \sum n_i u_i / \sum n_i$. Make an orthogonal linear transformation from u_i to new variables, one of which is a multiple of $\bar{u}$. The remainder will necessarily involve only contrasts of the u_i, and hence log-odds, and will be appropriate to the multinomial situation. Consequently the posterior distribution of the log-odds is proportional to $\exp\{-\frac{1}{2}\sum n_i(u_i - \bar{u})^2\}$ and, as in the argument leading to

(3.2), the posterior probability that $\sum n_i(u_i - \bar{u})^2 \leqq \chi_\alpha^2$ is $(1 - \alpha)$, the χ^2 having $(k - 1)$ degrees of freedom. The test of the null hypothesis will be significant at the α level if the null values for u_i do not belong to the confidence set. The null value for $u_i - \bar{u}$ is $(\ln \theta_i^{(0)} - \ln n_i) - \{\sum n_i(\ln \theta_i^{(0)} - \ln n_i)\}/\sum n_i$. If n_i/N replace n_i, as they may do without affecting the expression, we have here an expression involving comparisons of the logarithms of the *observed* proportions n_i/N, with the logarithms of the *expected* proportions, $\theta_i^{(0)}$. The difference between these is d_i, say, and the expression to be compared with χ_α^2 is the weighted sum of squares of the d_i about their weighted mean; the weights being n_i.

The test just derived replaces, in the present analysis, the usual test based on Pearson's expression $\sum (O - E)^2/E$: using instead the deviations $(\ln O - \ln E)$. Now the Pearson test may be extended to provide a test of a composite null hypothesis by replacing the expected numbers, E, under the simple null by the expected numbers under the maximum likelihood value for the composite null: the χ^2 criterion losing a degree of freedom for each parameter estimated by maximum likelihood. (Any other asymptotically efficient method may be used.) The same extension to χ^2 is similarly available in a Bayesian analysis and details are given in [15]. It is also possible to extend the test just derived to the composite case by replacing $\theta_i^{(0)}$ by any asymptotically efficient estimate, $\hat{\theta}_i$, of θ_i under the composite null hypothesis. The same test criterion is used with $\hat{\theta}_i$ replacing $\theta_i^{(0)}$ and the degrees of freedom reduced by one for each parameter estimated. We do not give the proof here since it closely parallels that for Pearson's criterion which is rather long. The interested reader with the above reference beside him should have no difficulty in making the necessary alterations to the log-odds case. An example of the use of this method is provided by a problem concerning linkage in genetics. The observational material consists of a multinomial with $k = 4$, under the null hypothesis the four probabilities are respectively $9 + \alpha$, $3 - \alpha$, $3 - \alpha$, and $1 + \alpha$, each divided by 16. There α is the linkage parameter and under the null may be asymptotically estimated by $\hat{\alpha}$, say. The deviations, $\ln\{(9 + \hat{\alpha})/16\} - \ln n_i/N$ etc., between log-expected and log-observed, are then found and the test criterion is the weighted sum of squares of these about their weighted mean. This can be referred to χ^2 on $(k - 1) - 1$, here 2, degrees of freedom. It is only necessary to use this method when the null hypothesis cannot be expressed in terms of log-odds. When it can, simpler methods are available as we shall see below.

One advantage of the present approach over the usual χ^2-analysis is that it provides the whole of the posterior distribution and not just that aspect of it provided by a significance test. Thus (4.2), when integrated with respect to $\bar{u}$, provides the posterior density of all the $(k - 1)$ log-odds. If any log-odds are of special interest then they may considered on their own, integrating out the unwanted log-odds from the joint posterior distribution. A simpler approach is to consider the required log-odds directly. For example, suppose one is interested in the conditional probability of an observation falling in the first class given that it has fallen in the first two; that is, $\theta_1/(\theta_1 + \theta_2)$. Denote this parameter by ϕ, then $\ln [\phi/(1 - \phi)] = \ln \theta_1 - \ln \theta_2$ is a log-odds with variance $n_1^{-1} + n_2^{-1}$. Hence $\ln [\phi/(1 - \phi)]$ has a normal posterior distribution with mean $\ln n_1 - \ln n_2$ and variance $n_1^{-1} + n_2^{-1}$. Note that the same result may be obtained by regarding $(n_1 + n_2)$ as fixed and treating the situation as binomial with ϕ as the parameter.

This argument does not immediately extend to the situation where the parameter cannot be expressed in terms of log-odds. An example is provided by $\theta_1 + \theta_2$. However the example may be handled by combining the first two classes, when $\theta_1 + \theta_2$ is simply the probability of falling into this new class, and inferences may be made using the results of Section 2. But this method involves a prior distribution over the parameters of the new $(k - 1)$ classes that needs to be shown to be equivalent to the original prior over the k classes. Otherwise one is changing the prior to suit the analysis which may be a good approximation but is not correct Bayesian argument. In fact, no approximation is involved and the two priors are equivalent. Thus, in order to make inferences about $\theta_1 + \theta_2$ the original multinomial over k classes can be replaced by a binomial over two. To establish the equivalence of the priors it is convenient to pass to contingency tables, which exhibit the combination of classes in its commonest form. In a contingency table of r rows and s columns, $k = rs$ and if one is interested in the row probabilities, these are the sum of the s probabilities in that row.

5. Two-way contingency tables. A two-way contingency table typically arises through each of a number of items being classified into one of r exclusive and exhaustive classes $A_1, A_2, \cdots A_r$, and simultaneously into one of s exclusive and exhaustive classes $B_1, B_2, \cdots B_s$. The probability θ_{ij} of an item being classified as both A_i and B_j is supposed the same for each item. If N items are independently classified, the numbers $n_{ij} (i = 1, 2, \cdots r; j = 1, 2, \cdots s)$ of items in classes A_i and B_j are multinomially distributed with index N and parameters θ_{ij}. The probability of being classified as A_i is $\sum_{j=1}^{s} \theta_{ij} = \theta_{i\cdot}$, say. $\theta_{\cdot j} = \sum_{i=1}^{r} \theta_{ij}$ is similarly the probability of being classified as B_j. In the usual language, the description just given refers to a contingency table with neither margin (only the total, N) fixed. We shall have something to say about contingency tables with one or both margins fixed below.

It is often convenient to parameterize the contingency table other than through the θ_{ij}. For example, the probabilities of the classes $A_i (i = 1, 2, \cdots r)$, $\theta_{i\cdot}$, and the conditional probabilities of the classes B_j, given the A_i, $\theta_{ij}/\theta_{i\cdot}$, may be used instead. Inferences about these cannot be made directly until their prior distribution has been determined. To this end we prove the following

THEOREM 2. *If the prior distribution of $\theta_{ij} (i = 1, 2, \cdots r: j = 1, 2, \cdots s)$ is proportional to $\prod_{ij} \theta_{ij}^{-1}$, then the prior distribution of $\theta_{i\cdot}$ and $\phi_{ij} = \theta_{ij}/\theta_{i\cdot}$ is proportional to*

$$\prod_{i} \theta_{i\cdot}^{-1} \prod_{i,j} \phi_{ij}^{-1}. \tag{5.1}$$

Consider the change from parameters θ_{ij}, for all i and j except $i = r, j = s$, to $\theta_{i\cdot} (i < r)$ and $\phi_{ij} (j < s)$. Certain values have to be excluded because of the constraints that exist amongst the parameters; namely $\sum_{ij}\theta_{ij} = \sum_i \theta_{i\cdot} = \sum_j \phi_{ij} = 1$. We have

$$\begin{aligned} &\theta_{ij} = \phi_{ij}\theta_{i\cdot} && i < r, j < s, \\ (5.2) \qquad &\theta_{rj} = \phi_{rj}(1 - \sum_{i<r} \theta_{i\cdot}) && j < s, \\ &\theta_{is} = (1 - \sum_{j<s} \phi_{ij})\theta_{i\cdot} && i < r. \end{aligned}$$

The Jacobian of the transformation from θ_{ij} to the new parameters is easily calculated to be

$$(5.3) \qquad \prod_{i=1}^{r} \theta_{i\cdot}^{s-1}.$$

Consequently the joint density of ϕ_{ij} and $\theta_{i\cdot}$ is proportional to the product of $\prod_{i,j}\theta_{ij}^{-1}$ and (5.3): that is to

$$\prod_{i=1}^{r}\prod_{j=1}^{s}(\theta_{i\cdot}\phi_{ij})^{-1}\prod_{i=1}^{r}\theta_{i\cdot}^{s-1} = \prod \theta_{i\cdot}^{-1}\prod \phi_{ij}^{-1},$$

as required.

The theorem establishes that the prior distribution on the rs classes is consistent with the same prior distribution on the reduced number r of classes. In particular the argument used at the end of the previous section is justified, and generally classes may be amalgamated and the approximation of Theorem 1 applied to the reduced number of parameters. Even more can be said because of properties of the likelihood function.

The multinomial likelihood for the contingency table is proportional to

$$\prod_{ij}\theta_{ij}^{n_{ij}} = \prod_i \theta_{i\cdot}^{n_{i\cdot}}\prod_{i,j}\phi_{ij}^{n_{ij}},$$

where $n_{i\cdot} = \sum_j n_{ij}$, a product of a function of the $\theta_{i\cdot}$ and a function of the ϕ_{ij}. The theorem proved that the prior similarly factors. The same is therefore true of the posterior distribution and inferences concerning $\theta_{i\cdot}$ can be made independently of those concerning ϕ_{ij}. Inferences about the marginal probabilities $\theta_{i\cdot}$ follow as for the multinomial on r classes. Inferences for $\phi_{ij}(j = 1, 2, \cdots s)$ for any i, similarly follow using a multinomial on s classes since the prior distribution $\prod_j \phi_{ij}^{-1}$ is of the usual form. There is a symmetrical result in terms of $\theta_{\cdot j}$ and $\theta_{ij}/\theta_{\cdot j}$ using the other margin based on the B-classification.

Another consequence of these results is that inferences about the conditional probabilities, ϕ_{ij}, can be made irrespective of the distribution of the $n_{i\cdot}$. The reason for this is that the only part of the likelihood used in the inference is the conditional probability of n_{ij} given $n_{i\cdot}$: the distribution of the $n_{i\cdot}$ is not involved. Contingency tables often arise with the $n_{i\cdot}$ fixed, rather than random variables, as when the number of items in each class A_i is selected deterministically and then classified according to the B-classification. The table is often said to have one margin fixed. Contingency tables with one margin fixed may be analyzed in the same way as tables with only N fixed; at least as far as the ϕ_{ij} are concerned, inferences about the $\theta_{i\cdot}$ are not possible in that case.

Let us apply the results to the 2×2 table ($r = s = 2$). A convenient parameterization is through $\theta_{1\cdot} = \theta_{11} + \theta_{12}$, the probability of the class A_1, and ϕ_{11} and ϕ_{21}, the probabilities of the classification B_1 given A_1 and A_2 respectively. Combining the results of the two theorems we see that the posterior distributions are approximately as follows:

(a) $\ln \{\theta_{1\cdot}/\theta_{2\cdot}\}$ is normal with mean $\ln \{n_{1\cdot}/n_{2\cdot}\}$ and variance $n_{1\cdot}^{-1} + n_{2\cdot}^{-1}$,
(b) $\ln \{\phi_{11}/\phi_{12}\}$ is normal with mean $\ln \{n_{11}/n_{12}\}$ and variance $n_{11}^{-1} + n_{12}^{-1}$,
(c) $\ln \{\phi_{21}/\phi_{22})$ is normal with mean $\ln \{n_{21}/n_{22}\}$ and variance $n_{21}^{-1} + n_{22}^{-1}$,

and these three distributions are independent. All three results follow because in each case we are dealing with a binomial situation and the three likelihoods and their corresponding priors factorize. Symmetrical results are available with A and B interchanged. Inference (a) is not available if the margin corresponding to the A-classification is fixed.

Usually a more interesting parameter is one describing the relationship between the two classifications. Indeed, much of the literature on 2×2 tables is only concerned with this aspect and even more specifically is devoted to examining whether the two classifications are independent: that is, $\theta_{ij} = \theta_{i\cdot}\theta_{\cdot j}$. Two tests available are Fisher's exact test and the usual χ^2-approximation. The latter has a Bayesian justification: no similar rationalization seems to exist for Fisher's method though progress might be possible on the lines suggested by Bahadur and described below. Measurement of the association between the classifications can be carried out in a number of ways: we seek for one in terms of log-odds. If the classifications are independent $\phi_{11} = p(B_1 \mid A_1) = \theta_{11}/(\theta_{11} + \theta_{12}) = \theta_{21}/(\theta_{21} + \theta_{22}) = p(B_1 \mid A_2) = \phi_{12}$ and this may be written either as

$$\theta_{11}/\theta_{12} = \theta_{21}/\theta_{22} \quad \text{or} \quad \phi_{11}/\phi_{12} = \phi_{21}/\phi_{22},$$

or in words, the odds for the B-classification are the same within A_1 and A_2. Hence a possible parameter to consider is the log-odds

$$(5.4) \qquad \begin{aligned} \phi &= \ln \theta_{11} - \ln \theta_{21} - \ln \theta_{12} + \ln \theta_{22}, \\ &= \ln \phi_{11} - \ln \phi_{21} - \ln \phi_{12} + \ln \phi_{22}. \end{aligned}$$

By the main theorem, or by a combination of (b) and (c) above, this parameter is approximately normally distributed with mean

$$(5.5) \qquad \ln n_{11} - \ln n_{21} - \ln n_{12} + \ln n_{22}$$

and variance

$$(5.6) \qquad n_{11}^{-1} + n_{21}^{-1} + n_{12}^{-1} + n_{22}^{-1}.$$

The null hypothesis of independence is $\phi = 0$ and may be tested by referring

$$(5.7) \qquad \frac{(\ln n_{11} - \ln n_{21} - \ln n_{12} + \ln n_{22})^2}{n_{11}^{-1} + n_{21}^{-1} + n_{12}^{-1} + n_{22}^{-1}}$$

to χ^2 on one degree of freedom. The approximation can probably be improved by subtracting $\frac{1}{2}$ from each of the n_{ij} in the numerator before taking logarithms. This result is available for tables with one or no margins fixed. If ϕ is used as a measure of association, Bayesian confidence intervals for it may easily be obtained using (5.5) and (5.6). The natural parameters to use with ϕ are $\theta_{1\cdot}$ and $\theta_{\cdot 1}$, the marginal probabilities.

It will be helpful in passing to tables larger than 2 × 2 to note certain similarities between the analysis and an analysis of variance. If the table (with no margins fixed) is analyzed in terms of $\theta_{1\cdot}$, $\theta_{\cdot 1}$ and ϕ, the first two correspond to main effects of the A- and B-classifications separately. The last corresponds to an interaction between the two classifications: indeed, the form of (5.4) is exactly that of an interaction based on the logarithms of the probabilities. However, if ϕ is thought of as an interaction a corresponding main effect would be

$$\ln \theta_{11} - \ln \theta_{21} + \ln \theta_{12} - \ln \theta_{22}, \tag{5.8}$$

which is usually of no interest. What is of interest, and what has been used, is ((a) above)

$$\ln \{\theta_{1\cdot}/\theta_{2\cdot}\} = \ln (\theta_{11} + \theta_{12}) - \ln (\theta_{21} + \theta_{22}). \tag{5.9}$$

Even had (5.8) been used, it should be noted that it is not independent of ϕ since the $\ln \theta_{ij}$ do not have equal variances. Whilst therefore analysis of variance ideas are conceptually useful in the study of contingency tables the breakdown of the total variation is into parts that are of separate interest and are not necessarily or typically independent: neither do the individual contributions have an additive property that is often demanded of analyses of variance or of χ^2. The interdependence of the posterior distributions of $\theta_{1\cdot}$, $\theta_{\cdot 1}$ and ϕ is as follows. $\theta_{1\cdot}$ is independent of ϕ ((a)—(c) above) as is $\theta_{\cdot 1}$. But $\theta_{1\cdot}$ is not independent of $\theta_{\cdot 1}$ nor is ϕ independent of the pair $(\theta_{1\cdot}, \theta_{\cdot 1})$. No parameterization that is symmetrical in the two classifications and gives independent distributions seems possible.

The 2 × 2 contingency table occasionally arises with both margins, $n_{i\cdot}$ and $n_{\cdot j}$, fixed. The classic example is the lady tasting tea. The above ideas do not seem to lead to a Bayesian solution partly because of the difficulty mentioned at the end of the last paragraph. Bahadur, in a private communication, has suggested a parameterization of the conditional distribution, given the two margins, that does lead to a Bayesian analysis. If the two classifications are independent it is easy to calculate this distribution for n_{11} : it is hypergeometric. Denote it by $p(n_{11})$. Then the suggestion is to use

$$p(n_{11} \mid \theta) = e^{\theta n_{11}} p(n_{11}) / \sum_n e^{\theta n} p(n) \tag{5.10}$$

for the distribution when the classifications are not independent. The parameter θ measures the amount of association between the two classifications: $\theta = 0$ corresponding to independence. The likelihood is sufficiently simple for Bayesian analysis to be possible but its value is limited by the difficulty in interpretation of θ. But this is perhaps a criticism of the design of the experiment because in any case it is difficult to see how from any analysis one could infer what the chance was of the lady classifying correctly any future cup of tea presented to her. The distributions (5.10) have been used by Elfving in a Neyman-Pearson study of the problem.

The extension of the analysis of the 2×2 table to the $r \times 2$ table for general r is straightforward. Inferences about $\theta_{\cdot 1}$ or about the $\theta_{i\cdot} (i = 1, 2, \cdots r)$ can be made in the ways described above for the binomial and multinomial respectively. The null hypothesis of no association between the two classifications is that $\theta_{ij}/\theta_{i\cdot} = \phi_{ij}$ does not depend on i. In terms of log-odds this is the same as saying

$$\ln \theta_{i1} - \ln \theta_{i2} \tag{5.11}$$

does not depend on i, or that all the log-odds in (5.11) are equal. Now these log-odds are independent and hence their posterior density has logarithm proportional to

$$\sum\{(\ln \theta_{i1} - \ln \theta_{i2}) - (\ln n_{i1} - \ln n_{i2})\}^2(n_{i1}^{-1} + n_{i2}^{-1})^{-1}. \tag{5.12}$$

This may be written (compare the passage from (4.1) to (4.2)) in terms of a weighted sum of squares about the mean and a term involving the mean. Consequently a test that all of (5.11) are equal can be made with the former. If $x_i = \ln n_{i1} - \ln n_{i2}$ and $m_i = n_{i1}^{-1} + n_{i2}^{-1}$ the test criterion is

$$\sum (x_i - x_\cdot)^2 m_i\,, \tag{5.13}$$

where $x_\cdot = \sum m_i x_i / \sum m_i$, and is referred to χ^2 on $(r - 1)$ degrees of freedom·

This test may be derived another way, which serves again to establish a relationship between these ideas and the analysis of variance. The hypothesis (5.11) is equivalent to saying that there exist constants a_i, b_j $(i = 1, 2, \cdots r; j = 1, 2)$ such that

$$\ln \theta_{ij} = a_i + b_j\,. \tag{5.14}$$

Furthermore, provided only contrasts are considered, the $\ln \theta_{ij}$ are independently normally distributed with known variances n_{ij}^{-1}. Consequently the hypothesis (5.14) is the hypothesis of no interaction in a two-way analysis of variance with the usual normal distribution theory, the variances being known but unequal. The roles of parameters and observations are interchanged: it is the parameters that are the random variables. The hypothesis may be tested by the usual method of considering the appropriate residual sum of squares when the model (5.14) is fitted. The details are given in Section 4.4 of [18]. That this method is available in the Bayesian framework is shown in detail in [15], but that it should be so is intuitively obvious from the results given in Section 3 of this paper. It is not difficult to demonstrate that the method of analysis of variance leads to the same result as (5.13). It may be noted that the approach based on (5.13) is simpler than Scheffé's approach when one of the classifications is dichotomous.

New difficulties arise when we pass to the general $r \times s$ table with both r and s greater than 2. It is not possible to find log-odds that are both independent and reflect the hypothesis of interest, namely that the two classifications are independent. It is easy to satisfy the latter requirement but without the former the quadratic form corresponding to (5.12) is no longer a simple sum of squares. Independence cannot be achieved since the variances of the $\ln \theta_{ij}$ are chance quantities. For example, consider a test of no association between the classifications in a 3×3 table. Four log-odds whose vanishing would be equivalent to the independence of the classifications are

$$\begin{aligned} &\ln \theta_{11} - \ln \theta_{12} - \ln \theta_{21} + \ln \theta_{22}, \\ &\ln \theta_{11} - \ln \theta_{13} - \ln \theta_{21} + \ln \theta_{23}, \\ &\ln \theta_{21} - \ln \theta_{22} - \ln \theta_{31} + \ln \theta_{32}, \\ &\ln \theta_{21} - \ln \theta_{23} - \ln \theta_{31} + \ln \theta_{33}. \end{aligned} \tag{5.15}$$

But these are correlated; for example, the covariance between the first and second is $n_{11}^{-1} + n_{21}^{-1}$. The direct way to proceed from here is to determine the dispersion matrix, A, of the log-odds in (5.15) and also the sample values of the same contrasts: namely the expressions that result from replacing θ_{ij} in (5.15) by n_{ij}. If n is the column vector of the sample values then the relevant quadratic form is $n'A^{-1}n$ which may be referred to χ^2 on four degrees of freedom. The same quadratic form would result whatever log-odds were chosen since they must be linear in the log-odds used in (5.15) and, as mentioned in Section 3, the method is invariant under linear transformations. The calculation is not too prohibitive on an electronic computer. If only the test is required, and not confidence limits for the log-odds, it is not even necessary to invert A, since if the triangular resolution $A = T'T$ is used, the quadratic form is the sum of squares $m'm$ where $T'm = n$.

Nevertheless the computations involve square matrices of size $(s - 1) \times (t - 1)$, and it may be better to proceed directly to the analysis of variance approach and calculate the criterion as a residual sum of squares. To do this it is necessary to solve linear equations in either $(r - 1)$ or $(s - 1)$ unknowns, whichever is the smaller. The matrices involved are therefore much smaller in size, but the manipulations on them are more complicated and the computations of the individual entries less simple. The direct calculations can be considerably simplified by using methods suggested by Goodman [9]. He shows that the inversion problems can in part be easily solved and that the final matrices that have to be inverted in any numerical case are substantially smaller than described above. For details we refer the reader to Goodman's paper.

6. Three-way contingency tables. We next consider the problem of $r \times s \times t$ tables in which individuals are classified in each of three ways: the two already referred to and also the exclusive and exhaustive classes $C_1, C_2, \cdots C_t$. Denote the numbers and probabilities in the classes by n_{ijk} and θ_{ijk} respectively. As in the two-way table it is convenient to reparameterize the situation and, for example, use

$$(6.1)\quad \theta_{i\cdot\cdot} = \sum_{j,k} \theta_{ijk}, \qquad \phi_{ij\cdot} = \sum_{k} \theta_{ijk}/\theta_{i\cdot\cdot}, \qquad \Psi_{ijk} = \theta_{ijk}/\sum_{k} \theta_{ijk}.$$

Then $\theta_{i\cdot\cdot}$ is the probability of A_i, ϕ_{ij} is the probability of B_j, given A_i, and Ψ_{ijk} is the probability of C_k given both A_i and B_j. Clearly $\theta_{ijk} = \theta_{i\cdot\cdot}\phi_{ij\cdot}\Psi_{ijk}$. Repeated applications of Theorem 2 show that if the prior density is proportional to $\prod \theta_{ijk}^{-1}$ then the posterior distributions of $\{\theta_{i\cdot\cdot}\}$, $\{\phi_{ij\cdot}\}$ and $\{\Psi_{ijk}\}$ will be independent. Furthermore, from the form of the likelihood function, the distribution of $\{\phi_{ij\cdot}\}$ will not depend on the distribution (if any) of $n_{i\cdot\cdot}$, nor will the distribution of $\{\Psi_{ijk}\}$ depend on the distribution (if any) of $n_{ij\cdot}$. These results enable tables in which one or two of the classifications are nonrandom to be analyzed. The methods already discussed enable analyses of $\{\theta_{i\cdot\cdot}\}$ and $\{\phi_{ij\cdot}\}$ to be made: we therefore consider $\{\Psi_{ijk}\}$.

The fact that the methods for two-way tables do extend to larger tables is, we claim, one of the main advantages of them. The classical methods based on the χ^2-statistic present complications when an attempt is made to extend them. One extension to 2 × 2 × 2 tables that is well known is Bartlett's [1] definition of no three-factor interaction. The hypothesis is that

$$(6.2)\qquad \frac{\theta_{111}\theta_{221}}{\theta_{211}\theta_{121}} = \frac{\theta_{112}\theta_{222}}{\theta_{212}\theta_{122}}.$$

This can equivalently be written with Ψ everywhere replacing θ, the suffixes remaining unaltered. If logarithms are taken of both sides of (6.2) we are led to consider

$$(6.3)\qquad \begin{aligned} \Psi = \{&\ln \theta_{111} - \ln \theta_{211} - \ln \theta_{121} + \ln \theta_{221}\} \\ &- \{\ln \theta_{112} - \ln \theta_{212} - \ln \theta_{122} + \ln \theta_{222}\}, \end{aligned}$$

which is a log-odds, and whose vanishing corresponds to no interaction in Bartlett's sense. Ψ is easily interpreted as the difference for classes C_1 and C_2 of the measure of association (5.4) between the A- and B-classifications. By the symmetry of (6.3) the letters A, B and C may be permuted in any way in the last sentence. The approximate posterior distribution of Ψ is normal with mean equal to the same expression with n_{ijk} for θ_{ijk}, and variance $\sum_{i,j,k} n_{ijk}^{-1}$.

The parameter Ψ can be used as a definition of the three-factor interaction in a 2 × 2 × 2 table. This claim is easily substantiated since (6.3) is exactly the form of such an interaction in the analysis of variance of $\ln \theta_{ijk}$. It is therefore possible, by the methods described, to investigate a 2 × 2 × 2 table using seven parameters: $\theta_{1\cdot\cdot}$, $\theta_{\cdot 1\cdot}$, $\theta_{\cdot\cdot 1}$; the three two-factor interactions corresponding to the measures of association defined in (5.4); and the three factor interaction Ψ. (The two-factor interactions can clearly be written in terms of $\phi_{ij\cdot}$, $\phi_{i\cdot k}$ and $\phi_{\cdot jk\cdot}$) This is a complete breakdown analogous to that in the analysis of variance but it should be noted that the parameters are not in general independent and that the additivity present in the standard analysis, or in the analyses by χ^2, Lancaster [13] or using information ideas, Kullback et al. [12] does not obtain here. If one or two of the classifications are non-random then only some of the parameters can be discussed.

The connection with the analysis of variance is an important one but there are often other parameters that are of interest besides the main effects and interactions already mentioned. For example, the above treatment is symmetric in the three classifications and yet it may be that the main purpose of the experiment that led to the contingency table was the investigation of the dependence of one classification on the others. If so, an unsymmetric analysis is clearly suggested. The Bayesian analysis provides, typically in an unmanageable form, the joint posterior distribution of all the parameters. What the statistician has to do is to extract from this material the main features that are likely to be of interest to the experimenter or of value in any decision problem for which the data are relevant. There must therefore be considerable flexibility in the type of analysis, or condensation of the posterior distribution that is used. Slavish attention to analysis of variance ideas is not likely to lead to many fruitful results. We now proceed to illustrate other approaches that are possible within the Bayesian log-odds framework, emphasizing that they are only illustrative and particular situations may demand other parameterizations. The discussion is within the framework of a 2 × 2 × 2 table. Extensions to larger three-way tables are indicated at the end of the section.

Suppose that it is the dependence of the C-classification on the other two that is of interest. Then we may think of A and B as providing two factors whose influence on the dependent variable, represented by the C-classification, we wish to assess. If A and B are non-random then the only probabilities that are defined are

$$\Psi_{ij1} = p(C_1 \mid A_i, B_j), \qquad (i, j = 1, 2) \tag{6.4}$$

with $\Psi_{ij2} = 1 - \Psi_{ij1}$. These four probabilities may be compared in various ways using the log-odds

$$\ln p(C_1 \mid A_i, B_j) - \ln p(C_2 \mid A_i, B_j) \tag{6.5}$$

and one comparison is the expression (6.3). One possibility is to investigate whether

$$p(C_1 \mid B_j, A_1) = p(C_1 \mid B_j, A_2) \qquad (j = 1, 2). \tag{6.6}$$

If this obtains then, given the B-classification, the C- and A-classifications are independent. Or alternatively, given B_j, A_i provides no further information about A. If this holds for both B_1 and B_2 then the A-classification provides no information not already provided by the B-classification. It is useful to refer to this as a *Markov* property, in analogy with the corresponding property studied in stochastic process theory where A, B and C refer to three points in time conveniently thought of as the past, present and future respectively.

The equalities (6.6) may be tested using the pair of log-odds

$$\begin{aligned}(6.7)\qquad &\{\ln p(C_1 \mid B_j, A_1) - \ln p(C_2 \mid B_j, A_1)\} \\ &\qquad - \{\ln p(C_1 \mid B_j, A_2) - \ln p(C_2 \mid B_j, A_2)\} (j = 1, 2)\end{aligned}$$

with means

$$(6.8)\qquad \ln n_{1j1} - \ln n_{1j2} - \ln n_{2j1} + \ln n_{2j2},$$

and variance

$$(6.9)\qquad n_{1j1}^{-1} + n_{1j2}^{-1} + n_{2j1}^{-1} + n_{2j2}^{-1}.$$

The two expressions in (6.7) for $j = 1$ and 2 are independent and hence the Markov property may be tested by referring the sum for $j = 1$ and 2 of the squares of the means, (6.8), divided by the variances, (6.9) to χ^2 on 2 degrees of freedom. (The null value of (6.7) is zero.) If the posterior distribution is sufficiently concentrated around zero then one might feel reasonably confident that one could proceed on the basis that the Markov property obtained. If so, then it is possible to discuss the simpler conditional probability $p(C_1 \mid B_j)$, which is otherwise undefined. It is not easy to see exactly what is meant by "sufficiently concentrated" in the sentence above. The degree of departure from the Markov property that is allowable will depend on the "robustness" of the analysis that results from using the property, to departures from it. This is not a subject that can be discussed here: some considerations of it in a different context are given in Lindley [14].

Consider next the case where it is still the dependence of C on A and B that is of interest, but where one of the independent classifications, A say, is random and the other, B, is not. Then, in addition to the conditional probabilities $p(C_k \mid A_i, B_j)$ already discussed, the probabilities $p(C_k \mid B_j)$ are also meaningful. The analysis of variance approach can be misleading in this context because the natural main effects and interactions are rather differently defined in contingency table analyses from the usual linear hypothesis situation. (The same point has already been discussed in the two-way table; compare equations (5.8) and (5.9).) The main effect of B on C is defined in terms of an average over the levels of A: in fact

$$(6.10)\qquad p(C_k \mid B_j) = p(C_k \mid A_1, B_j)p(A_1 \mid B_j) + p(C_k \mid A_2, B_j)p(A_2 \mid B_j)$$

which is defined since the A-classification is random. Since $p(A_1 \mid B_j) \neq p(A_2 \mid B_j)$ in general this is a weighted average of the probabilities for each A-classification. The "factors" A and B are therefore "confounded" and the main effect is difficult to interpret. An example, due to Simpson [19], also quoted by Edwards [5] will clarify the issue. The data are as follows

(6.11)

	Male		Female	
	Untreated	Treated	Untreated	Treated
Alive	4	8	2	12
Dead	3	5	3	15.

The dependent C-classification is into Alive or Dead. The random independent A-classification is by sex, and the non-random independent B-classification is by treatment. Let us ignore sampling variations and suppose the proportions (out of a total of 52 persons) are the probabilities $\{\theta_{ijk}\}$. The 2×2 table for Males shows that the association between treatment and death is $-5/6$ when measured by the difference of log-odds (Equation (5.4)). The same value is obtained for the 2×2 table for Females. Consequently, as measured by the difference of log-odds the treatment has been equally beneficial for the two sexes. This is equivalent to saying that the three-factor interaction (6.3) is zero. But if the two tables for Males and Females are combined to produce results that do not refer to sex, we obtain

(6.12)

	Untreated	Treated
Alive	6	20
Dead	6	20

and the association between treatment and death is zero: or the two classifications are independent. The explanation is that the allocation of treatments and sex are confounded, as can be seen by considering the 2×2 table that results from ignoring the dependent classification into Alive or Dead. The result is

(6.13)

	Male	Female
Untreated	7	5
Treated	13	27

.

The females, with their higher death rate have been treated more often than the males. Consequently it is not easy to understand the separate effects of treatment and sex. In the usual analysis of variance situations it is possible to separate the effects to some extent but this method is not available with contingency tables since the association in (6.12) is not a linear function of those in the separate sex tables in (6.11). In (6.12) we use $\ln \theta_{\cdot jk}$ and not linear forms in $\ln \theta_{1jk}$ and $\ln \theta_{2jk}$ as in linear hypothesis situations (compare again equations (5.8) and (5.9)). Had the proportions of each sex undergoing treatment been the same at 40/52 the table corresponding to (6.12) would have had about the same measure of association as the separate 2×2 tables in (6.12). As it is, the only course is to consider the sexes separately.

The extension to general $r \times s \times t$ tables is straightforward. The interaction of all three factors may be defined in terms of the parameters

$$(6.14) \quad \begin{aligned}&\{\ln \theta_{ijk} - \ln \theta_{rjk} - \ln \theta_{isk} + \ln \theta_{rsk}\}\\ &\qquad - \{\ln \theta_{ijt} - \ln \theta_{rjt} - \ln \theta_{ist} + \ln \theta_{rst}\}\end{aligned}$$

for $i < r, j < s, k < t$. (Compare Equations (5.15).) If these are all zero then there is no interaction. The vanishing of (6.4) may be tested in the usual way, though it must be noticed that the log-odds in (6.14) for different values of i, j and k are not independent and the dispersion matrix needs to be found. The details of the calculations are substantially reduced by using the ideas of Goodman [9].

The Markov property that given B, the A- and C-classifications are independent may be based on consideration of the log-odds

$$\{\ln p(C_k \mid B_j, A_i) - \ln p(C_t \mid B_j, A_i)\} - \{\ln p(C_k \mid B_j, A_r) - \ln p(C_t \mid B_j, A_r)\} \tag{6.15}$$

for $i < r$, $k < t$ and all j. For different j these are independent. For each j there are $(r - 1)(t - 1)$ log-odds which are not independent, but the quadratic form yields a χ^2-statistic having $(r - 1)(t - 1)$ degrees of freedom. The total number of degrees of freedom is therefore $(r - 1)s(t - 1)$.

7. Relationship with previous work. There is little earlier work known to the author on analysis of contingency tables that uses a prior distribution: exceptions are Jeffreys [10], in particular Section 5.11; and two papers by Good [6], [7]. Nevertheless much of the material that falls within the classical framework is relevant and illuminating for a Bayesian. It is well known and almost obvious that if one is dealing with a normal distribution of unknown mean and known variance, or generally with any location parameter, where the density is of the form $f(x - \theta)$ for the random variable x and the parameter θ, there is a close parallelism between the classical analyses based on x for fixed θ and the Bayesian analysis using the likelihood function of θ for fixed x. As an example consider statements like (3.2) and (3.3). The same is broadly true in the multinomial context and corresponding to our Theorem 1 there is a result that the distribution of $\ln n_i$ for fixed θ_i is approximately normal with mean $\ln N\theta_i$ and variance $(N\theta_i)^{-1}$, and they are independent provided linear contrasts in the $\ln n_i$ are considered. This result has been used by Plackett [16] to provide classical analyses of contingency tables along lines closely similar to ours. In particular he has considered the problem of testing an interaction, a problem which had previously been studied by Roy and Kastenbaum [17]. Goodman [9] has shown how Plackett's computations can be simplified in the way outlined at the end of Section 5. The first mention of the method based on the property of $\ln n_i$ known to me is Woolf's use of it for the $2 \times 2 \times t$ table [20]. A paper by Darroch [4] is relevant in deciding which hypotheses might be considered. Good [8] has provided a general definition of interaction which is equivalent to ours in the case of a 2^n table but does not seem so easy to handle for tables involving classifications into more than two exclusive and exhaustive classes. Good has resolved a conjective made by Darroch. Birch [3] has discussed definitions of interactions similar to those proposed here.

The methods advocated by Lancaster [13] based on the breakdown of χ^2 suffer from several defects. There is no justification for the use of the statistic, nor for the breakdown into additive components. The methods have been criticized by Plackett.

The methods of Kullback [11] and [12] do not appear to have a Bayesian interpretation despite the use of information-theoretic concepts. Again the breakdown into additive components is unnatural in many situations.

It is relevant to remark that the measure of association suggested for a 2 × 2 table, Equation (5.4), has been shown to have certain attractive properties by Edwards [5]. Specifically any measure of association which is a function of the conditional probabilities $p(A_i \mid B_j)$ and equally of the conditional probabilities $p(B_j \mid A_i)$ is necessarily a function of ϕ.

8. Future extensions. The prior distribution used throughout this paper is the special one proportional to $\prod \theta_i^{-1}$. But as explained in the first section, the results are available whenever the prior distribution is equivalent, in the amount of information it contains, to data obtained from a contingency table of the same type as that to be analyzed, with the special prior. The two tables, hypothetical and real, may be combined and the analysis based on the combined table and the special prior.

It would be highly desirable to extend the analyses to more general priors. For example, one might have a 2 × 2 table with little prior knowledge of $\theta_{i\cdot}$ but substantial prior knowledge of ϕ_{ij} (for the notation see the beginning of Section 5). Such knowledge might be equivalent to few observations on the margin of the 2 × 2 table but many observations on the interior of the table for a selected set of marginal totals. Another example is a trinomial situation with classes A_1, A_2 and A_3. One may feel fairly sure that θ_1 is around 0.20 whereas knowledge of θ_2 and θ_3 (apart from the fact that they add up to 0.80) is slight.

Another type of prior that needs consideration is one that allows correlations between the $\ln \theta_i$. All the analyses in the present paper hinge on independence of them. For example in a 2 × 2 table one may have prior knowledge that the two probabilities $p(A_1 \mid B_j)(j = 1, 2)$ are close in value, and it would be desirable to incorporate this into the analysis. Indeed, there are situations where this may be essential: as where B_1 corresponds to the "at-homes" and B_2 to the "not-at-homes" in a social survey. A second example of a correlated prior is where the multinomial distribution has arisen from a grouped frequency distribution: that is, a histogram. The smoothness of the underlying density produces a correlation between neighboring groups.

It is hoped to show in a future paper that such situations can be handled using log-odds and multivariate normal distributions.

REFERENCES

[1] Bartlett, M. S. (1935). Contingency table interactions. *J. Roy. Statist. Soc. Suppl.* **2** 248–252.

[2] Bartlett, M. S. and Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *J. Roy. Statist. Soc. Suppl.* **8** 128–138.

[3] Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **25** 220–233.

[4] Darroch, J. N. (1962). Interactions in multi-factor contingency tables. *J. Roy. Statist. Soc. Ser. B* **24** 251–263.

[5] Edwards, A. W. F. (1963). The measure of association for 2 × 2 tables. *J. Roy. Statist. Soc. Ser. A* **126** 109–113.

[6] Good, I. J. (1957). Saddle-point methods for the multinomial distributions. *Ann. Math. Statist.* **28** 861–881.

[7] Good, I. J. (1956). On the estimation of small frequencies in contingency tables. *J. Roy. Statist. Soc. Ser. B* **18** 113–124.

[8] Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

[9] Goodman, L. A. (1963). On Plackett's test for contingency table interactions. *J. Roy. Statist. Soc. Ser. B* **25** 179–188.

[10] Jeffreys, H. (1961). *Theory of Probability.* Oxford: Clarendon Press.

[11] Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York.

[12] Kullback, S., Kupperman, M. and Ku, H. H. (1962). Tests for contingency tables and Markov chains. *Technometrics* **4** 573–608.

[13] Lancaster, H. O. (1951). Complex contingency tables treated by the partition of χ^2. *J. Roy. Statist. Soc. Ser. B* **13** 242–249.

[14] Lindley, D. V. (1961). The robustness of interval estimates. *Bull. Inst. Internat. Stat.* **38** 209–220.

[15] Lindley, D. V. (1964). *Introduction to Probability and Statistics.* Cambridge Univ. Press.

[16] Plackett, R. L. (1962). A note on interactions in contingency tables. *J. Roy. Statist. Soc. Ser. B* **24** 162–166.

[17] Roy, S. N. and Kastenbaum, M. A. (1956). On the hypothesis of no 'interaction' in a multi-way contingency table. *Ann. Math. Statist.* **27** 749–757.

[18] Scheffé, H. (1959). *Analysis of Variance.* Wiley, New York.

[19] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.

[20] Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann. Human Genetics* **19** 251–253.