Study Factors at More Than Two Levels

The preceding discussion on the analysis of retrospective data has been in terms of the test factor under study taking only two values. This framework has sufficed for discussion of the underlying statistical ideas and issues. In practice, the study factor will frequently take on more than two, perhaps many, potential values. When the number of study factor values is large, grouping can reduce them to manageable proportions.

The need to consider only a limited number of classes for the study factor stems from the fact that, when an association is anticipated, most of the significant information about the association will come from the results for the more extreme values of the study factor. While it is efficient to concentrate attention on the test factor classes expected to show the greatest differences in association with the disease, it is also profitable to consider intermediate values for the test factor to seek evidence for a consistent pattern of association. For example, in Table 1, a highly significant difference between nonsmokers and women currently smoking more than 1 pack of cigarettes daily was illustrated. Inclusion of data for smokers of 1 pack or less a day showing results intermediate between the other classes would have added little, if anything, to the statistical significance of the results, and might actually lower it, if one made an over-all test of the differences among the three smoking classes. However, the observation that the intermediate smoking class does, in fact, show an intermediate relative risk contributes to an orderly pattern and increases our confidence in the conclusions suggested by the data for the remaining two classes.

For any two particular test-factor levels, the relative risk for one over the other may be calculated using only the data pertaining to those two levels or by using the results for all test levels. In the formulas previously given for \( R, R_1, R_2, R_3, \) and \( R_4 \), the difference between the two calculating procedures is simply one of setting the values of \( N_{1i}, N_{2i}, \) and \( T_i = N_{1i} + N_{2i} \) in terms of number of cases and controls occurring at the two study-factor levels only, or defining them in terms of total number of cases and controls in the entire study. When total cases and controls are used in defining \( N_{1i}, N_{2i}, \) and \( T_0 \), it can be shown that for \( R, R_1, R_2, \) and \( R_3 \) the various relative risks will be internally consistent with each other. If the relative risk for the first level is twice that for the second level, which in turn is twice that for the third level, then the relative risk for the first level will be four times that of the third. These exact relationships do not hold for \( R \) as an estimator of relative risk, and a somewhat sophisticated extension of the formula for \( R \) would be required to secure this property.

The problem of obtaining a summary chi square when the study factor is at more than two levels is complicated by the fact that the deviations from expectation at the various study-factor levels are intercorrelated. When there are but two levels, the two deviations will have perfect negative correlation, and attention need be directed to only one of the deviations. Irrespective of the number of levels, at any one level the deviation from expectation among diseased persons will be equal, but opposite in sign, to the deviation from expectation among controls, so that attention can be confined to the deviations for diseased persons.

The problem can be stated as one of reducing a set of correlated deviations into a summary chi square. Table 2 applies this process for obtaining a summary chi square to the study of the association of epidermoid and undifferentiated pulmonary carcinoma in women and maximum cigarette-smoking rate, classified into three levels, after adjustment for age and occupation.

The general expressions for the expectations and variances of the number of cases at a particular test-factor level are given in the lower right section of Table 2. Also shown is the expression for the covariance between the number of cases at two different test-factor levels. Since the total of all the deviations is zero, one would in general need the variances of, and covariances between, the number of cases at all but one of the levels. The number of covariance terms will rise sharply as the number of test levels are increased. At 3 test levels, there are 2 variance terms and 1 covariance term, while at 10 test levels, there would be 9 variances and 36 covariance terms of interest.

For the general case the burden of computation could be heavy. After all the necessary computation for the deviations, their variances and covariances, there would still remain the problem of converting these, presumably by matrix methods, into a summary chi square. Since the retrospective problem will normally involve only a limited number of test-factor levels, precise procedures will be given only for the three-level situation, and approximate procedures outlined for the general case.

The exact computation procedure for the three-level case is detailed in Table 2. Lines (1), (2), and (4) show the total observed and expected frequencies and variances of the number of cases (and controls) at each of the three smoking-rate levels, after adjusting for age and occupation. These are the summary totals over each subclassification obtained by application of the formulas appearing in Table 2.

Lines (5) and (6) give the chi squares corresponding to the total deviation from expectation at each of the smoking-rate levels. The chi squares in line (5) are corrected for continuity. They relate to the difference of the particular level to which they apply, from the two other levels combined. Following the usual practice of making no continuity corrections when chi squares with more than 1 degree of freedom are under consideration, line (6) shows the uncorrected chi squares.

The computing procedure of Table 2 takes advantage of the fact that, since the sum of the deviations from expectation is zero, the variance of the third deviation must equal the sum of the other two variances plus twice the covariance for the first two deviations. The covariance of the first two deviations is readily obtained as illustrated and is used in calculating the summary chi square. The summary chi square is obtained as the sum of squares of two orthogonal deviates, with each
Table 2—Illustrative computations of summary chi square, when there are 3 levels for the test factor and 2 levels for the second factor

<table>
<thead>
<tr>
<th>Category</th>
<th>Total (b)</th>
<th>Category (c)</th>
<th>Total (b)</th>
<th>Category (c)</th>
<th>Total (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epidermoid carcinoma</td>
<td>102</td>
<td>Occasional or non-smokers</td>
<td>302</td>
<td></td>
<td>302</td>
</tr>
<tr>
<td>Unidentified cancer</td>
<td>350</td>
<td></td>
<td>350</td>
<td></td>
<td>350</td>
</tr>
<tr>
<td>Total</td>
<td>452</td>
<td></td>
<td>452</td>
<td></td>
<td>452</td>
</tr>
</tbody>
</table>

For the generalization of this table to the case of \( \phi \) levels for the first and \( \psi \) levels for the second factors, the following formulas apply:

\[
\chi^2 = \sum_{i=1}^{\phi} \sum_{j=1}^{\psi} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

where \( O_{ij} \) is the observed frequency and \( E_{ij} \) is the expected frequency for the \( i \)-th level of the first factor and \( j \)-th level of the second factor.

The expected frequency for the \( i \)-th level of the first factor and \( j \)-th level of the second factor is given by:

\[
E_{ij} = \frac{R_i C_j}{N}
\]

where \( R_i \) is the row total for the \( i \)-th level of the first factor, \( C_j \) is the column total for the \( j \)-th level of the second factor, and \( N \) is the total number of observations.

In the present instance the summary chi square with 2 degrees of freedom is 28.43 [line (11)]. This presumably is close to the chi square with 1 degree of freedom which would have obtained had only the two most extreme smoking classes been compared. If one examines the individual uncorrected chi squares [line (6)], their total is found to be 45.55, the maximum individual figure being 23.42. It will necessarily be true that the summary chi-square value will lie between the largest of the three chi squares and their total. At almost any reasonable probability level these limits would be sufficient to establish statistical significance without further calculation. In our companion paper (97) this rule sufficed in almost all instances to separate the significant from the nonsignificant results.

Comments on Extensions to More Than Three Factors

Two procedures can be suggested for getting approximate summary chi squares, when there are a large number of levels for the test factors, without the burden of computation that the exact method would entail. Both methods calculate the approximate summary chi square as a sum of squares of approximately orthogonal standardized deviates.

In the first method one computes an uncorrected chi square with 1 degree of freedom for the difference of the first level from all the remaining levels combined (the same first step as in the illustration for the three-level case). Discarding the data from the first level, a second chi square is computed for the difference between the second test-factor level and the remaining levels combined. This is done successively up to and including the last two remaining levels. The approximate summary chi square is then the sum of the separate chi squares with the number of degrees of freedom being one less than the number of test levels.

Exactly orthogonal standardized deviates would be obtained if, in the summary analysis, as each successive total deviation from expectation were evaluated, it was adjusted for its multiple regression on the preceding deviations, and then standardized by the adjusted variance. This, of course, would no longer be a simplified approximate procedure. However, it can be shown that for a single classification, in the multiple regression of any deviation from expectation on any subset of deviations, the regression coefficients will all be equal; the multiple regression on the set of deviations will be the same as the simple regression on their sum. The equality of regression coefficients, while holding true exactly for deviations in the separate subclassifications, will hold only approximately for the total.

square adjusted for its own variance. The first deviate squared is simply the uncorrected chi square at the first level in line (6)—the variance of the deviate remaining as initially calculated. The second deviate is the deviation at the second level adjusted for its correlation with the first deviation [adjusted \( Y_2 = Y_2 - b_{21} Y_1 \); \( b_{21} = \text{covariance}(Y_1, Y_2)/\text{variance}(Y_1) \)]. The variance of the adjusted second deviate is the initial value reduced by that portion of the variation accounted for by the first deviation [Var. (adjusted \( Y_2 \)) = variance \( Y_2 - \text{covariance}(Y_1, Y_2)/\text{variance}(Y_1) \)].
deviations from expectation (it would hold exactly if equal numbers of individuals were observed from level to level at each subclassification). Nevertheless, this result suggests that approximately orthogonal deviations would be obtained if, in evaluating each successive total deviation, it were adjusted for the cumulative total of deviations already evaluated. Computing procedures to accomplish this can readily be devised.

Both approximate chi-square procedures just outlined, which may have merit when more than three groups are being compared simultaneously, should, in theory, yield linear combinations of independent chi squares. While testing the chi-square values obtained as though they were exact is not likely to be too inexact, it may be more correct to obtain a modified number of degrees of freedom, along the lines suggested by Satterthwaite (47) for problems involving such linear combinations. What the modified number of degrees of freedom would be has not been investigated by us, and it may prove as easy to apply the exact chi-square procedure, indicated later, as to determine the appropriate degrees of freedom for the approximate chi square.

It is of interest that a somewhat similar task of obtaining an appropriate summary chi square appears in the birth-order problems described by Halperin (48). There, it was necessary to compare a set of total observations (across family sizes) with a set of total expectations, one for each birth order. Halperin described a matrix-inversion procedure for reducing the set of correlated deviations into a summary chi square. In that problem it can be shown that all the regression coefficients are equal in the multiple regression of the deviation at a particular birth order on the set of deviations at all succeeding birth orders. The second approximate method described previously for the present problem could thus be used exactly for the birth-order problem, permitting simplified computation of chi square. The procedure indicated by Halperin has the advantage of generality and could be applied to the current and related problems, if one obtained all the necessary variances and covariances and inverted the resulting matrix.

References


