Statistical analysis of individually matched case-control studies in epidemiology: factor under study a discrete variable taking multiple values

M. C. PIKE* AND J. CASAGRANDE*

Departments of Community Medicine, Pediatrics, and Pathology, University of Southern California School of Medicine, Los Angeles, California 90033, USA

AND

P. G. SMITH

Department of Health and Social Security, Cancer Epidemiology and Clinical Trials Unit, Department of the Regius Professor of Medicine, Oxford University

Pike, M. C., Casagrande, J., and Smith, P. G. (1975). British Journal of Preventive and Social Medicine, 29, 196-201. Statistical analysis of individually matched case-control studies in epidemiology: factor under study a discrete variable taking multiple values. It is now well established in case-control studies with individual matching that the statistical analysis must take the fact of individual matching into account. The problems of so doing have been thoroughly dealt with for the situation of the factor of interest being a two-level factor (all-or-none response) and with one or more matched controls per patient. These results are extended in this paper to the situation where the factor can assume multiple levels in studies where each patient is matched with a single control.

A method which is frequently used when selecting a control group in epidemiological studies is to choose a control subject for each case, the control being 'matched' with the case for factors which are not of direct interest in the investigation but which are known, or suspected, to be associated with the disease. In the absence of matching, these factors might confound comparisons between the case and control groups when examining the specific factor under investigation. Such individual matching is, for example, often undertaken for variables such as age, sex, social class, and area of residence.

The statistical evaluation of such investigations must take account of this matching procedure (see, for example, McNemar, 1947; Mantel and Haenszel, 1959; Billewicz, 1964; Miettinen, 1969; Pike and Morrow, 1970). As well as testing the statistical significance of observed differences in the factor of interest in case-control pairs, it is usually necessary to obtain an estimate of the relative risk (r.r.) associated with the factor under study.

In the most simple situation, in which n controls have been matched with n cases and the factor under study is either present or not (or can take only two values), the results of the investigation may be tabulated as in Table I. The maximum likelihood estimate, r, of the true r.r., r, associated with the factor is given by the ratio \( n_{21}/n_{12} \) (not \( (n_{21}+n_{22})/(n_{11}+n_{21}) \)) \( (n_{11}+n_{21})/(n_{11}+n_{12}) \) \( (n_{12}+n_{22}) \), which is the usual estimate of r.r. in studies in which there is no individual case-control matching; and the statistical significance level is calculated from the standard \( \chi^2 \) test for equality of \( n_{12} \) and \( n_{21} \), namely:

\[
\begin{array}{ccc}
\text{Control} & \text{Absent} & \text{Present} \\
\text{Case} & n_{11} & n_{12} & n_{11} + n_{12} \\
\text{Absent} & n_{21} & n_{22} & n_{21} + n_{22} \\
\text{Present} & n_{11} + n_{21} & n_{12} + n_{22} & n \\
\end{array}
\]
where 
\[ \chi_1^2 = \left[ n_{21} - E(n_{21}) \right] - \frac{1}{2} V(n_{21}) \]

\[ E(n_{21}) = \text{expected value of } n_{21} = (n_{21} + n_{12})/2 \]

and 
\[ V(n_{21}) = \text{variance of } n_{21} = (n_{21} + n_{12})/4 \]

(for details see Pike and Morrow, 1970).

In analysing a recent case-control study we were faced with a more general situation in which the factor under study could take more than two values. The problem may be illustrated by the fictitious data on the relationship between the ABO blood groups and a particular disease shown in Table II.

### Table II

<table>
<thead>
<tr>
<th>Blood Group</th>
<th>Control</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>O</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>74</td>
<td>14</td>
<td>6</td>
<td>6</td>
<td>66</td>
<td>160</td>
</tr>
<tr>
<td>Case</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>AB</td>
<td>10</td>
<td>12</td>
<td>2</td>
<td>12</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>O</td>
<td>18</td>
<td>8</td>
<td>3</td>
<td>64</td>
<td>93</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
<td>38</td>
<td>13</td>
<td>146</td>
<td>301</td>
<td></td>
</tr>
</tbody>
</table>

The r.r. of disease for a particular blood group, say A, compared with another group, say AB, may ‘crudely’ be estimated as above, that is, \( \tilde{r}_{A,AB} = 6/10 = 0.60 \). In general it will be found that these \( \tilde{r} \) estimates are not mutually consistent. From the above table, for example, we find

\[ \tilde{r}_{A,B} = 14/2 = 7.00 \]
\[ \tilde{r}_{A,O} = 66/18 = 3.67 \]
\[ \tilde{r}_{B,O} = 4/8 = 0.50 \]

so that \( \tilde{r}_{A,B} \neq \tilde{r}_{A,O}/\tilde{r}_{B,O} \): whereas \( r_{A,B} \) must equal \( r_{A,O}/r_{B,O} \), that is the true r.r. of A compared with B must equal the true r.r. of A to O divided by the true r.r. of B to O.

This is clearly an undesirable property of these crude r.r. estimates and the problem of obtaining mutually consistent estimates in this situation of one to one matching and a multiple value factor is dealt with in this paper.

### Maximum Likelihood Approach

Suppose that there are \( k \) possible levels \( (A_1, A_2, \ldots, A_k) \) of the factor \( A \) under study. The results of the case-control investigation may be represented as in Table III.

#### Table III

<table>
<thead>
<tr>
<th>Factor Level</th>
<th>( A_1 )</th>
<th>( A_i )</th>
<th>( A_j )</th>
<th>( A_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A_1 )</td>
<td>( n_{11} )</td>
<td>( n_{1i} )</td>
<td>( n_{1j} )</td>
<td>( n_{1k} )</td>
</tr>
<tr>
<td>( A_i )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A_j )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( A_k )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n_j = number of case-control pairs in which the case was at level ( A_i ) and the control at level ( A_j )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Let \( r_{ji} = r.r. \) of \( A_j \) compared with \( A_i \). (If \( A_i \) and \( A_j \) were the only factor levels of interest, then the maximum likelihood estimate of \( r_{ji} \) would be simply \( n_{ji}/n_{ij} \).) Define

\[ r_j = r_{jk} \]

where

\[ r_k = 1 \]

then

\[ r_{ji} = r_{j}/r_i \]

The likelihood of the \( n_{ij} \)'s conditional on the sums \( (n_{ij} + n_{ji}) = N_{ij} + N_{ji} \) is proportional to

\[ l = \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} r_{ji} n_{ij} (1 + r_{ji}) N_{ij} \]

The log-likelihood function expressed in terms of the \( r_{ij} \)'s is therefore, except for a constant, equal to

\[ L_r = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_{ij} \log (r_{ij}) - \log \left( \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_{ij} \right) \]

The first derivative of \( L_r \) with respect to \( r_h \) is

\[ \delta L_r/\delta r_h = \sum_{i=1}^{h-1} (n_{hi}/r_h - N_{hi}/(r_i + r_h)) \]

\[ - \sum_{j=h+1}^{k} (n_{jh}/r_{h} - N_{jh}/(r_{h}^2 + r_h r_{j})) \]

The second derivative of \( L_r \) with respect to \( r_h \) is

\[ \frac{\delta^2 L_r}{\delta r_h^2} = \sum_{i=1}^{h-1} \frac{n_{hi}}{r_h^2 (r_i + r_h)^2} \]

\[ - \sum_{j=h+1}^{k} \frac{n_{jh}}{(r_{h}^2 + r_h r_{j})^2} \]
ESTIMATION

Maximum likelihood estimates \( \hat{r}_i \)'s of the \( r_i \)'s \((i=1, 2, \ldots, k-1)\) may be obtained by setting (4) to zero for \( h=1, 2, \ldots, k-1 \) simultaneously.

TESTS OF SIGNIFICANCE

Approximate significance levels may be derived using likelihood ratio tests.

We may first do an overall test for evidence of the \( r_i \)'s departing from unity. If we assume that

\[
r_1 = r_2 = \ldots = r_{k-1} = 1
\]

then the log-likelihood, ignoring the constant term, is given by

\[
L_1 = -\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_{ij} \log(2)
\]

Thus, if (5) holds,

\[
C_1 = -2(L^* \log(L_1))
\]

will be distributed approximately as \( \chi^2 \) with \((k-1)\) degrees of freedom, where \( L^* \) is the value of (3) when \( r_i \) is set equal to \( \hat{r}_i \) \((i=1, 2, \ldots, k-1)\).

CONFIDENCE INTERVALS

The log-likelihood function \( L_r \) may be far from symmetric in the \( r_i \)'s around its maximum and the derivation of approximate confidence intervals for the \( r_i \)'s is better accomplished working through the transformations

\[
t_i = r_i/(1 + r_i)
\]

(for \( k = 2 \), \( t_1 \) is the binomial parameter for \( n_{12} \) and \( n_{21} \)).

Expressing (3) in terms of the \( t_i \)'s we have

\[
L_t = \Sigma \Sigma n_{ij} \log \left(t_j - \Sigma \Sigma n_{ij} \log \left(1 - t_j \right) - \Sigma \Sigma n_{ji} \log \left(t_i + \Sigma \Sigma n_{ji} \log \left(1 - t_i \right) - \Sigma \Sigma N_{ij} \log \left(1 + t_j(1 - t_i)/t_i(1 - t_j) \right) \right) \right)
\]

and

\[
\delta L_t / \delta t_h = \sum_{i=1}^{h-1} \left\{ n_{hi}/t_h + n_{hi}/(1-t_h) \right\}
- N_{hi} (1-t_i)/(s_{ih}(1-t_h)) \right\}
- \sum_{j=h+1}^{k} \left\{ n_{hj}/t_h + n_{hj}/(1-t_h) \right\}
- N_{hj}t_j/(s_{jh}t_h)
\]

where \( s_{ij} = t_i(1-t_i) + t_j(1-t_j) \).

Thus

\[
E(n_{ij}) = N_{ij} t_i(1-t_j)/s_{ij}
\]

The variance-covariance matrix associated with the estimates \( t_i \) \((i=1, 2, \ldots, k-1)\) may be obtained by inverting the matrix of the expectations of the second partial differentials, which has diagonal elements (14) and off diagonal elements (15). Thus, by assuming approximate normality, confidence limits may be assigned to the \( t_i \)'s and hence to the \( r_i \)'s. In particular, in this way, the statistical significance of any one of the estimates \( \hat{r}_i \) may be assessed.

We have thus assigned confidence intervals to the estimates \( \hat{r}_i \) which are equal to \( \hat{r}_{ik} \). To determine the confidence interval associated with \( \hat{r}_{ji} = \hat{r}_j/\hat{r}_i \), it is simplest to redefine the rows and columns of Table III so that row \( i \) and row \( k \) are interchanged (and columns \( i \) and \( k \) are simultaneously reversed) and hence \( r_j \) will be redefined as \( r_{ij} \). Confidence intervals may then be determined as above by inversion of the revised matrix of expectations of second differentials to give the variance-covariance matrix associated with the new \( r_j \)'s.

(For completeness the second derivatives of \( L_r \) with respect to the \( r_i \)'s are given in Appendix I.)
Testing for Consistency

The $r_i$'s have been defined in such a way that mutually consistent estimates of $r.r.$ may be derived. However, it is generally desirable to test the reasonableness of this imposed restriction. In general there are $k(k-1)/2$ $r.r.$'s associated with the given problem. These are $r_{ji}$ ($i = 1, 2, \ldots, k-1; j = i+1, i+2, \ldots, k$) and the 'unrestricted-inconsistent' maximum likelihood estimates are

$$r_{ji} = n_{ji}/n_{ij}$$

(16)

The log-likelihood, ignoring the constant term, is given by

$$L_2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} n_{ji} \log(r_{ji}) - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_{ji}(1 + r_{ji})$$

and, if the assumption of consistent relative risks is correct, then

$$C_2 = -2(L^*_2 - L^*_1)$$

will be distributed approximately as $\chi^2$ with $(k-1)(k-2)/2$ degrees of freedom, where $L^*_2$ is the value of (17) when $r_{ji}$ is set equal to $r_{ji}$ ($i = 1, 2, \ldots, k-1; j = i+1, i+2, \ldots, k$) and $L^*_1$ is the value of (3) when $r_i$ is set equal to $\hat{r}_i$ ($i = 1, 2, \ldots, k-1$).

An Example

For the data given in Table II let $A_1 = $ blood group A, $A_2 = $ B, $A_3 = $ AB, and $A_4 = $ O. Then

$$\hat{r}_{1,4} = 66/18 = 3.667,$$

$$\hat{r}_{2,4} = 4/8 = 0.500,$$

$$\hat{r}_{3,4} = 12/3 = 4.000$$

The maximum likelihood consistent estimates are computed as (FORTRAN IV program available from authors on request):

$$\hat{r}_1 = 3.503, \hat{r}_2 = 0.559, \hat{r}_3 = 4.669$$

with

$$L^*_1 = -81.40$$

$$L_1 = -108.82$$

so that the overall test of significance is

$$C_1 = 54.84$$

distributed approximately as $\chi^2_3 (p < 0.0001)$.

$$L_2 = -81.14$$

so that the test for consistency assumption is

$$C_2 = 0.52 \sim \chi^2_2 (p \approx 0.90).$$

The $t_i$'s are

$$t_1 = 0.778, t_2 = 0.358, t_3 = 0.824$$

The approximate variance-covariance matrix of the $t_i$'s is

$$0.001747 \quad 0.001052 \quad 0.000894$$

$$0.001052 \quad 0.0009824 \quad 0.001582$$

$$0.000894 \quad 0.001582 \quad 0.003415$$

so that, for example, the approximate 95% confidence interval for $t_2$ is

$$t_2 \pm 1.96 \sqrt{(0.0009824)}$$

$$= 0.164, 0.552$$

Thus the approximate 95% confidence interval for $r_2$ is $(0.196, 1.232)$.

An illustrative example for the case $k = 3$, which may be computed on a desk calculator, is shown in Appendix II.

Discussion

The problem discussed in this paper is peculiar to studies in which there is individual matching of cases and controls, and where the matching is 'relevant', that is, affects the probabilities of the different answers.

This problem would arise, for example, in an investigation of whether there is any association between a specific disease and contraceptive practices. We might choose controls individually matched for factors such as age, parity, social class, etc. and record the type of contraceptive most used in the previous 12 months, the type being classified as, say, none, oral contraceptive, intrauterine device or other methods. The method of analysis we propose ensures that all of the relevant data are used in estimating the relative risks, and to do this we only have to impose the restriction that the estimates derived are mutually consistent. The method leads to a direct statistical test of this assumption, although, being based on multiple degrees of freedom, the test is not very powerful. The assumption of consistency is implicit in the commonly used, but incorrect, method of analysis which ignores the matching in deriving the relative risk estimates.

Finding that the relative risk estimates, using the paired data, are not consistent would suggest that the data should be examined for interactions between the matching variables and the relative risks. For example, if contraceptive practice was associated with the disease under study in one group but not another (say, parous or non-parous), then the relative risks might be consistent but different in the two
groups. If the data are tabulated ignoring parity, then the resulting relative risks based upon the pooled data would not be consistent.

Requests for reprints: M. C. Pike, Edmondson Research Building, USC Medical School, 1840 N. Soto Street, Los Angeles, California 90032 USA.

REFERENCES
MCNEMAR, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153.

APPENDIX I
The second derivatives of $L_r$ are

$$
\begin{align*}
\delta^2 L_r/\delta r_h^2 &= \sum_{i=1}^{h-1} \left\{ -n_{hi}/r_h^2 + N_{hl}/(r_l+r_h) \right\} \\
- \sum_{j=h+1}^{k} \left\{ -n_{jh}/r_h^2 + N_{jh}/(2r_h+r_j)/(r_h+r_j) \right\}
\end{align*}
$$

$$
\delta^2 L_r/\delta r_h \delta r_g = N_{hg}/(r_g+r_h)^2
$$

Now

$$
E(n_{ij}) = N_{ij} r_i / (r_l + r_j)
$$

Thus

$$
E(\delta^2 L_r/\delta r_h^2) = - \sum_{i=1}^{h} \frac{N_{hr} r_i}{r_h (r_h + r_i)^2}
$$

$$
E(\delta^2 L_r/\delta r_h \delta r_g) = N_{hg}/(r_g+r_h)^2
$$

APPENDIX II
ILLUSTRATIVE COMPUTATION OF RELATIVE RISK ESTIMATES FOR RESPONSE VARIABLE TAKING THREE VALUES

Although the estimation procedures presented in this paper are most suitably performed on a computer, the simple case involving a response variable taking three possible values, may be relatively easily handled on a desk calculator. These computations are illustrated below.

Suppose our study yields a table of the form:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>27</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Control response</td>
<td>10</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>14</td>
<td>22</td>
</tr>
</tbody>
</table>

Set:

$$
\begin{align*}
a &= n_{13} - n_{21} = 2 \\
b &= n_{21} + n_{31} = 27 \\
c &= n_{31} - n_{12} = 3 \\
d &= n_{12} + n_{13} = 26 \\
0 &= n_{12} - n_{23} = 2
\end{align*}
$$

Then the equations (4) reduce to:

$$
\begin{align*}
\hat{r}_1 &= \frac{r_2 (e + f \hat{r}_2)}{(g - h \hat{r}_2)} \\
\hat{r}_2 &= \frac{\hat{r}_1 (a - b \hat{r}_1)}{(c \hat{r}_1 - d)}
\end{align*}
$$

Where $\hat{r}_1$ and $\hat{r}_2$ are the maximum likelihood estimates of relative risk.

The solution of (A2.1) and (A2.2) reduces to finding the real root of the cubic equation:

$$
\hat{r}_2^3 + A\hat{r}_2^2 + B\hat{r}_2 + C = 0
$$

where, if $D = f(bf - ch) = 20832$

then $A = [h (af - dh) + f (eb + gc) + e (fb - ch)]/D = 4656/20832$

$B = [e (ah + be) + g (ce + dh) + g (dh - af)]/D = 3600/20832$

$C = - g (dg + ae)/D = -12672/20832$

(A2.3) may be simply solved using the Newton-Raphson iterative method. We first 'guess' a value for $\hat{r}_2$—say $x$ (in general $x = n_{23}/n_{32}$ would seem reasonable), then an improved estimate is given by

$$
x - (x^3 + Ax^2 + Bx + C) / (3x^2 + 2Ax + B)
$$

This procedure may be repeated until an estimate of the required accuracy is obtained. In the example, we thus obtain successive values of $x$ of 0.857, 0.736, 0.718, 0.717, 0.717 and we may thus set $\hat{r}_2 = 0.717$ and, by substitution in (A2.1) $\hat{r}_1 = 0.827$
Statistical analysis of individually matched case-control studies

From equation (3)
\[ L_r^* = (n_{13} + n_{12}) \log (\hat{r}_1) + (n_{21} + n_{23}) \log (\hat{r}_2) \]
\[ - N_{12} \log (\hat{r}_1 + \hat{r}_2) - N_{13} \log (1 + \hat{r}_1) \]
\[ - N_{23} \log (1 + \hat{r}_2) = -54 \cdot 2153 \]
(note: \( N_{ij} = n_{ij} + n_{ji} \))

From equation (6)
\[ L_1 = -(N_{12} + N_{13} + N_{23}) \log (2) = -54 \cdot 7586 \]

From equation (17)
\[ L_r^* = n_{21} \log (n_{21}/n_{12}) + n_{31} \log (n_{31}/n_{13}) \]
\[ + n_{32} \log (n_{32}/n_{23}) - N_{21} \log (N_{21}/n_{12}) \]
\[ - N_{31} \log (N_{31}/n_{13}) - N_{32} \log (N_{32}/n_{23}) \]
\[ = -53 \cdot 9135 \]

Thus, to test whether \( r_1 \) and \( r_2 \) differ from unity, from (7) we compute
\[ C_1 = -2 (L_r^* - L_1) = 1 \cdot 09 \]
which may be tested as \( \chi^2 \) with 2 degrees of freedom.
The test of consistency, from (18), is
\[ C_2 = -2 (L_r^* - L_r^*) = 0 \cdot 60 \]
which may be tested as \( \chi^2 \) with 1 degree of freedom.

To obtain the variances and covariance note from (8) that
\[ \hat{t}_1 = \hat{r}_1 / (1 + \hat{r}_1) = 0 \cdot 453 \]
\[ \hat{t}_2 = \hat{r}_2 / (1 + \hat{r}_2) = 0 \cdot 418 \]

Let \( s = t_1 (1 - t_2) + t_2 (1 - t_1) = 0 \cdot 492 \)

Then the \( 2 \times 2 \) matrix, say \( M \), defined by (14) and (15) is such that
\[ M_{11} = (t_2 (1 - t_2) N_{21} + s^2 N_{31}) / (s^2 t_1 (1 - t_1)) = 214 \cdot 375 \]
\[ M_{22} = (t_1 (1 - t_1) N_{12} + s^2 N_{23}) / (s^2 t_2 (1 - t_2)) = 207 \cdot 862 \]
\[ M_{12} = M_{21} = -N_{12}/s^2 = -99 \cdot 147 \]

Let \( K = M_{11} M_{12} - M_{12} M_{21} = 34730 \cdot 289 \)

Then \( \text{Var} (t_1) = M_{22}/K = 0 \cdot 005985 \)
\[ \text{Var} (t_2) = M_{11}/K = 0 \cdot 006173 \]
\[ \text{Cov} (t_1, t_2) = -M_{12}/K = 0 \cdot 002855 \]

and approximate 95\% confidence limits on \( t_1 \) and \( t_2 \) are:
\[ t_1 = 0 \cdot 453 \pm 1 \cdot 96 \times 0 \cdot 077 = (0 \cdot 302, 0 \cdot 604) \]
\[ t_2 = 0 \cdot 418 \pm 1 \cdot 96 \times 0 \cdot 079 = (0 \cdot 263, 0 \cdot 573) \]

and thus approximate 95\% confidence limits on \( r_1 \) and \( r_2 \) are:
\[ r_1 = (0 \cdot 433, 1 \cdot 525) \]
\[ r_2 = (0 \cdot 357, 1 \cdot 342) \]

[noting that \( r_1 = t_1/(1 - t_1); r_2 = t_2/(1 - t_2) \)]