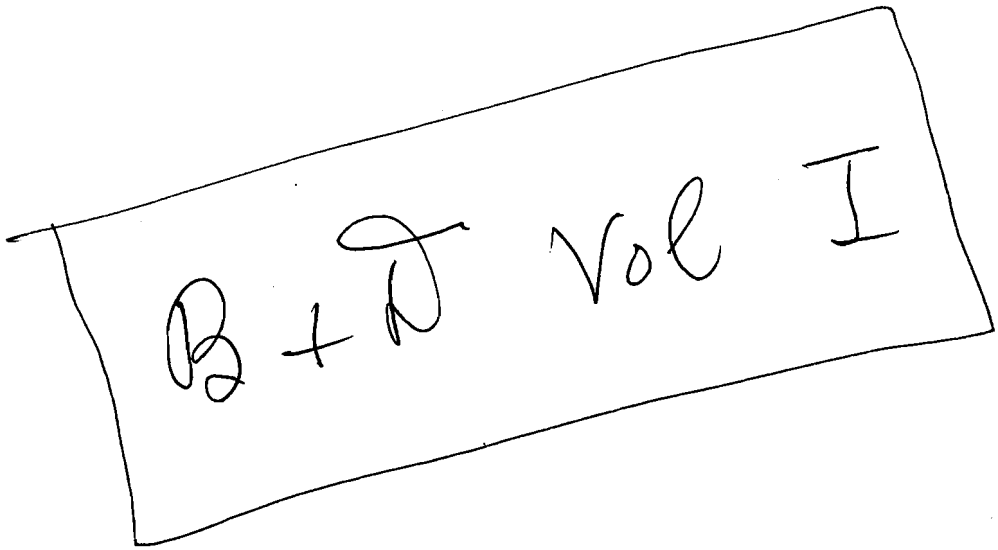


4. CLASSICAL METHODS OF ANALYSIS OF GROUPED DATA

CHAPTER IV

CLASSICAL METHODS OF ANALYSIS OF GROUPED DATA

This chapter presents the traditional methods of analysis of case-control studies based on a grouping or cross-classification of the data. The main outlines of this approach, which has proved extremely useful for practising epidemiologists, are contained in the now classic paper by Mantel and Haenszel (1959). Most of the required calculations are elementary and easily performed on a pocket calculator, especially one that is equipped with log, exponential and square-root keys. Rothman and Boice (1979) have recently published a set of programmes for such a calculator which facilitates many of the analyses presented below.

The statistical procedures introduced in this chapter are appropriate for study designs in which stratification or "group-matching" is used to balance the case and control samples *vis-à-vis* the confounding variables. The following chapter develops these same methods of analysis for use with designs in which there is individual matching of cases to controls. While our treatment of this material attempts to give the reader some appreciation of its logical foundations, the emphasis is on methodological aspects rather than statistical theory. Some of the more technical details are labelled as such and can be passed over on a first reading. These are developed fully in the excellent review papers by Gart (1971, 1979) on statistical inferences connected with the odds ratio. Fleiss (1973) presents an elementary and very readable account of many of the same topics. Properties of the binomial, normal and other statistical distributions mentioned in this chapter may be found in any introductory text, for example Armitage (1971).

4.1 The Ille-et-Vilaine study of oesophageal cancer

Throughout this chapter we will illustrate the various statistical procedures developed by applying them to a set of data collected by Tuyns et al. (1977) in the French department of Ille-et-Vilaine (Brittany). Cases in this study were 200 males diagnosed with oesophageal cancer in one of the regional hospitals between January 1972 and April 1974. Controls were a sample of 778 adult males drawn from electoral lists in each commune, of whom 775 provided sufficient data for analysis. Both types of subject were administered a detailed dietary interview which contained questions about their consumption of tobacco and of various alcoholic beverages in addition to those about foods. The analyses below refer exclusively to the role of the two factors, alcohol and tobacco, in defining risk categories for oesophageal cancer.

Table 4.1 summarizes the relevant data. Since no attempt had been made to stratify the control sample, there is a tendency for the controls to be younger than the cases,

Table 4.1 Distribution of risk factors for cases and controls: Ille-et-Vilaine study of oesophageal cancer[a]

	Cases	Controls
Age (years)		
25–34	1	115
35–44	9	190
45–54	46	167
55–64	76	166
65–74	55	106
75+	13	31
Mean	60.0	50.2
S.D	9.2	14.3
Alcohol (g/day)		
0–39	29	386
40–79	75	280
80–119	51	87
120+	45	22
Mean	84.9	44.4
S.D	48.4	31.9
Tobacco (g/day)		
0–9	78	447
10–19	58	178
20–29	33	99
30+	31	51
Mean	16.7	10.5
S.D	12.9	11.9

[a] Data taken from Tuyns et al. (1977)

Table 4.2 Correlations between risk variables in the control sample: Ille-et-Vilaine study of oesophageal cancer[a]

	Age	Tobacco	Alcohol
Age	1.0	–0.02	–0.02
Tobacco		1.0	0.15
Alcohol			1.0

[a] Data taken from Tuyns et al. (1977)

a feature which has to be accounted for in the analysis. Cases evidently have a history of heavier consumption of both alcohol and tobacco than do members of the general population. Correlations among these risk variables in the population controls indicate that there are no systematic linear trends of increased or decreased consumption with age, and that the two risk variables are themselves only weakly associated (Table 4.2).

To take an initial look at the relationship between alcohol and risk using traditional epidemiological methods, we might dichotomize alcohol consumption, using as a cut-off point the value (80 g/day) closest to the median for the cases, as there are many more controls. This yields the basic 2×2 table:

	Average daily alcohol consumption		
	80+ g	0–79 g	Total
Cases	96	104	200
Controls	109	666	775
Total	205	770	975

Of course such a simple dichotomization into "exposed" and "unexposed" categories can obscure important information about a risk variable, particularly concerning dose-response (§ 3.2). Summarizing the entire set of data in a single table ignores the possible confounding effects of age and tobacco (§ 3.4); these two deficiencies are momentarily ignored.

From the data in this or any similar 2×2 table one wants to estimate the relative risk and also to assess the degree of uncertainty inherent in that estimate. We need to know at what level of significance we can exclude the null hypothesis that the true relative risk ψ is equal to unity, and we need to determine a range of values for ψ which are consistent with the observed data. Appropriate methods for making such tests and estimates are presented in the next two sections.

4.2 Exact statistical inference for a single 2×2 table[1]

When a single stratum of study subjects is classified by two levels of exposure to a particular risk factor, as in the preceding example, the data may be summarized in the ubiquitous 2×2 table:

	Exposed	Unexposed	
Diseased	a	b	n_1
Disease-free	c	d	n_0
	m_1	m_0	N

(4.1)

A full understanding of the methods of analysis of such data, and their rationale, requires that the reader be acquainted with some of the fundamental principles of statistical inference. We thus use this simplest possible problem as an opportunity to review the basic concepts which underlie our formulae for statistical tests, estimates and confidence intervals.

Inferential statistics and analyses, as opposed to simple data summaries, attempt not only to describe the results of a study in as precise a fashion as possible but also to assess the degree of uncertainty inherent in the conclusions. The starting point for

[1] Parts of this section are somewhat technical and specialized; they may be skimmed over at first reading.

such analyses is a *statistical model* for the observed data which contains one or more *unknown parameters.* Two such models for the 2×2 table were introduced implicitly in the discussion in § 2.8. According to the first, which can be called the cohort model, the marginal totals of m_1 exposed and m_0 unexposed persons are regarded as fixed numbers determined by the sample size requirements of the study design. There are two unknown parameters, the probabilities P_1 and P_0 of developing the disease during the study period. Since subjects are assumed to be sampled at random from the exposed and unexposed subpopulations, the *sampling distribution* of the data is thus the product of two *binomial* distributions with parameters (P_1,m_1) and (P_0,m_0). In the second model, which is more appropriate for case-control studies, the marginal totals n_1 and n_0 are regarded as fixed by design. The distribution of the data is again a product of two binomials, but this time the parameters are (p_1,n_1) and (p_0,n_0) where p_1 and p_0 are the exposure probabilities for cases and controls.

According to § 2.8 the key parameter for case-control studies is the odds ratio ψ, partly because it takes on the same value whether calculated from the exposure or the disease probabilities. The fact that the probability distributions of the full data depend on two parameters, either (P_1,P_0) or (p_1,p_0), complicates the drawing of conclusions about the one parameter in which we are interested. Hypotheses that specify particular values for the odds ratio, for example, the hypothesis $H_0: \psi = 1$ of no association between exposure and disease, do not completely determine the distribution of the data. Statistics which could be used to test this hypothesis depend in distribution on *nuisance* parameters, in this case the baseline disease or exposure probabilities. Inferences are much simpler if we can find another probability distribution, using perhaps only part of the data, which depends exclusively on the single parameter of interest.

A distribution which satisfies this requirement is the *conditional distribution of the data assuming all the marginal totals are fixed.* Cox (1970) and Cox and Hinkley (1974) discuss several abstract principles which support the use of this distribution. Its most important property from our viewpoint is that the (conditional) probability of observing a given set of data is the same whether one regards those data as having arisen from a case-control or a cohort study. In other words, the particular sampling scheme which was used does not affect our inferences about ψ. Regardless of which of the two product binomial models one starts with, the probability of observing the data (4.1) conditional on all the marginal totals n_1,n_0,m_1,m_0 remaining fixed is

$$\text{pr}(a \mid n_1,n_0,m_1,m_0;\psi) = \frac{\binom{n_1}{a}\binom{n_0}{m_1-a}\psi^a}{\sum_u \binom{n_1}{u}\binom{n_0}{m_1-u}\psi^u} \tag{4.2}$$

Here $\binom{n}{u}$ denotes the *binomial coefficient*

$$\binom{n}{u} = \frac{n(n-1)(n-2)\ldots(n-u+1)}{u(u-1)(u-2)\ldots(1)}$$

which arises in the binomial sampling distribution. The summation in the denominator is understood to range over all values u for the number of exposed cases (a) which are possible given the configuration of marginal totals, namely $0, m_1-n_0 \leqq u \leqq m_1,n_1$.

Two aspects of the conditional probability formula (4.2) are worthy of note. First, we have expressed the distribution solely in terms of the number a of exposed cases. This is adequate since knowledge of a, together with the marginal totals, determines the entire 2×2 table. Expressing the distribution in terms of any of the other entries (b, c or d) leads to the same formula either for ψ or for ψ^{-1}. Second, the formula remains the same upon interchanging the roles of n and m, which confirms that it arises from either cohort or case-control sampling schemes.

As an example, consider the data

2	1	3	
3	1	4	(4.3)
5	2	7	

The possible values for a determined by the margins are a = 1, 2 and 3, corresponding to the two tables

1	2	3		3	0	3	
4	0	4	and	2	2	4	
5	2	7		5	2	7	

in addition to that shown in (4.3). Thus the probability (4.2) for a = 2 may be written

$$\frac{\binom{3}{2}\binom{4}{3}\psi^2}{\binom{3}{1}\binom{4}{4}\psi + \binom{3}{2}\binom{4}{3}\psi^2 + \binom{3}{3}\binom{4}{2}\psi^3}$$

$$= \frac{\frac{3\times 2}{2\times 1}\times\frac{4\times 3\times 2}{3\times 2\times 1}\psi^2}{\frac{3}{1}\times\frac{4\times 3\times 2\times 1}{4\times 3\times 2\times 1}\psi + \frac{3\times 2}{2\times 1}\times\frac{4\times 3\times 2}{3\times 2\times 1}\psi^2 + \frac{3\times 2\times 1}{3\times 2\times 1}\times\frac{4\times 3}{2\times 1}\psi^3}$$

$$= \frac{12\psi^2}{3\psi + 12\psi^2 + 6\psi^3} = \frac{4\psi}{1 + 4\psi + 2\psi^2}.$$

Similarly, the probabilities of the values a = 1 and a = 3 are $1/(1+4\psi+2\psi^2)$ and $2\psi^2/(1+4\psi+2\psi^2)$, respectively.

Estimation of ψ

The distribution (4.2) for $\psi \neq 1$ is known in the probability literature as the *non-central hypergeometric distribution.* When $\psi = 1$ the formula becomes considerably simpler and may be written

$$\text{pr}(a \mid n_1,n_0,m_1,m_0; \psi = 1) = \frac{\binom{n_1}{a}\binom{n_0}{m_1-a}}{\binom{n_1+n_0}{m_1}}, \tag{4.4}$$

which is the (central) hypergeometric distribution. So called *exact* inferences about the odds ratio ψ are based directly on these conditional distributions. The *conditional maximum likelihood estimate* $\hat{\psi}_{\text{cond}}$, i.e., the value which maximizes (4.2), is given by the solution to the equation

$$a = E(a \mid n_1,n_0,m_1,m_0; \psi), \tag{4.5}$$

where E denotes the expectation of the discrete distribution. For example, with the data (4.3) one must solve

$$2 = \frac{1+8\psi+6\psi^2}{1+4\psi+2\psi^2},$$

a quadratic equation with roots $\pm\sqrt{\frac{1}{2}}$, of which the positive solution is the one required. Note that this estimate, $\hat{\psi}_{\text{cond}} = \sqrt{\frac{1}{2}} = 0.707$, differs slightly from the *empirical odds ratio* $\frac{ad}{bc} = \frac{2}{3} = 0.667$. Unfortunately, if the data are at all extensive, (4.5) defines a polynomial equation of high degree which can only be solved by numerical methods.

Tests of significance

Tests of the hypothesis that ψ takes on a particular value, say $H:\psi = \psi_0$, are obtained in terms of *tail probabilities* of the distribution (4.2). Suppose, for example, that $\psi_0 = 10$. The conditional probabilities associated with each of the three possible values for a are then:

a	pr(a | 3,4,5,2; $\psi = 10$)
1	$\frac{1}{1+40+200} = 0.004$
2	$\frac{40}{1+40+200} = 0.166$
3	$\frac{200}{1+40+200} = 0.830$

Having observed the data (4.3), in which a = 2, the *lower* tail probability, $p_L = 0.004 + 0.166 = 0.17$, measures the degree of evidence against the hypothesis $H:\psi = 10$ in favour of the alternative hypothesis that $\psi<10$. While the data certainly suggest that $\psi<10$, the fact that p_L exceeds the conventional significance levels of 0.01 or 0.05 means that the evidence against H is weak. Much stronger evidence would be provided if a = 1, in which case the *p-value* or *attained significance level* is 0.004.

More generally, the lower tail probability based on the distribution (4.2) is defined by

$$p_L = \sum_{u \leqq a} \mathrm{pr}(u \mid n_1, n_0, m_1, m_0; \psi_0) \tag{4.6}$$

and measures the degree of evidence against the hypothesis $H:\psi = \psi_0$ in favour of $\psi<\psi_0$. Similarly, the upper tail probability

$$p_U = \sum_{u \geqq a} \mathrm{pr}(u \mid n_1, n_0, m_1, m_0; \psi_0) \tag{4.7}$$

measures the degree of evidence against H and in favour of $\psi>\psi_0$. In both cases the summation is over values of u consistent with the observed marginal totals, with u less than or equal to the observed a in (4.6) and greater than or equal to a in (4.7). If no alternative hypothesis has been specified in advance of the analysis, meaning we concede the possibility of a putative "risk factor" having either a protective or deleterious effect, it is common practice to report twice the minimum value of p_L and p_U as the attained significance level of a *two-sided* test[1].

The hypothesis most often tested is the *null* hypothesis $H_0:\psi = 1$, meaning no association between risk factor and disease. In this case the tail probabilities may be computed relatively simply from the (central) hypergeometric distribution (4.4). The resulting test is known as *Fisher's exact test.* For the data in (4.3) the exact upper p-value is thus

$$\left\{\binom{3}{2}\binom{4}{3}+\binom{3}{3}\binom{4}{2}\right\} \div \binom{7}{5} = 18/21 = 0.86,$$

while the lower p-value is

$$\left\{\binom{3}{1}\binom{4}{4}+\binom{3}{2}\binom{4}{3}\right\} \div \binom{7}{5} = 15/21 = 0.71,$$

neither of them, of course, being significant.

Confidence intervals

Confidence intervals for ψ are obtained by a type of testing in reverse. Included in the two-sided interval with a specified *confidence coefficient* of $100(1-\alpha)\%$ are all values ψ_0 which are *consistent with the data* in the sense that the two-sided p-value

[1] An alternative procedure for computing two-sided p-values is to add $p_{min} = \min(p_L, p_U)$ to the probability in the opposite tail of the distribution obtained by including as many values of the statistic as possible without exceeding p_{min}. This yields a somewhat lower two-sided p-value than simply doubling p_{min}, especially if the discrete probability distribution is concentrated on only a few values.

for the test of $H:\psi = \psi_0$ exceeds α. In other words, the confidence interval contains those ψ_0 such that both p_L and p_U exceed $\alpha/2$, where p_L and p_U depend on ψ_0 as in (4.6) and (4.7). In practice, the interval is determined by two endpoints, a *lower confidence limit* ψ_L and an *upper confidence limit* ψ_U. The upper limit satisfies the equation

$$\alpha/2 = \sum_{u \leqq a} \text{pr}(u \mid n_1, n_0, m_1, m_0; \psi_U) \tag{4.8}$$

while the lower limit satisfies

$$\alpha/2 = \sum_{u \geqq a} \text{pr}(u \mid n_1, n_0, m_1, m_0; \psi_L). \tag{4.9}$$

Thus the exact upper $100(1-\alpha)\% = 80\%$ confidence limit for the data (4.3) is obtained from the equation

$$\alpha/2 = 0.10 = \frac{1+4\psi}{1+4\psi+2\psi^2},$$

with solution $\psi_U = 18.25$, while the lower limit solves

$$\alpha/2 = 0.10 = \frac{4\psi+2\psi^2}{1+4\psi+2\psi^2},$$

with solution $\psi_L = 0.0274$. Since there are so few data in this example, most reasonable values of ψ are consistent with them and the interval is consequently very wide.

Although such exact calculations are feasible with small 2×2 tables like (4.3), as soon as the data become more extensive they are not. The equations for conditional maximum likelihood estimation and confidence limits all require numerical methods of solution which are not possible with pocket calculators. Thomas (1971) provides an algorithm which enables the calculations to be carried out on a high-speed computer, but extensive data will render even this approach impracticable. Fortunately, with such data, the exact methods are not necessary. We next show how approximations may be obtained for the estimates, tests and confidence intervals described in this section which are more than accurate enough for most practical situations. Occasionally the exact methods, and particularly Fisher's exact test, are useful for resolving any doubts caused by the small numbers which might arise, for example, when dealing with a very rare exposure. The exact procedures are also important when dealing with matched or finely stratified samples, as we shall see in Chapters 5 and 7.

4.3 Approximate statistical inference for a 2×2 table

The starting point for approximate methods of statistical inference is the *normal approximation* to the conditional distribution. When all four cell frequencies are large, the probabilities (4.2) are approximately equal to those of a continuous normal distribution whose mean $A = A(\psi)$ is the value which a must take on in order to give an empirical or calculated odds ratio of ψ (Hannan & Harkness, 1963). In other words, to find the *asymptotic mean* we must find a number A such that when A replaces a, and the remaining table entries are filled in by subtraction

A	$B = n_1 - A$	n_1
$C = m_1 - A$	$D = n_0 - m_1 + A$	n_0
m_1	m_0	N

, (4.10)

we have

$$\frac{AD}{BC} = \frac{A(n_0 - m_1 + A)}{(n_1 - A)(m_1 - A)} = \psi. \tag{4.11}$$

This is a quadratic equation, only one of whose roots yields a possible value for A in the sense that A, B, C and D are all positive. Under the special null hypothesis $H_0 : \psi = 1$, the equation simplifies and we calculate

$$A(1) = \frac{m_1 n_1}{N}, \tag{4.12}$$

which is also the mean of the exact distribution (4.4). The quantities A, B, C and D in (4.10) are known as *fitted values* for the data under the hypothesized ψ.

Once A is found and the table is completed as in (4.10), the *variance* $\mathrm{Var} = \mathrm{Var}(a;\psi)$ of the approximating normal distribution is defined in terms of the reciprocals of the fitted values

$$\mathrm{Var} = \left[\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}\right]^{-1}. \tag{4.13}$$

When $\psi = 1$ this reduces to

$$\mathrm{Var}\,(a; \psi = 1) = \frac{n_1 n_0 m_1 m_0}{N^3},$$

whereas the variance of the corresponding exact distribution is slightly larger

$$\mathrm{Var}(a; \psi = 1) = \frac{n_1 n_0 m_1 m_0}{N^2 (N-1)}. \tag{4.14}$$

Using the approximating normal distribution in place of (4.2) leads to computationally feasible solutions to the problems outlined earlier.

Estimation

The asymptotic maximum likelihood estimate is obtained by substituting the asymptotic mean $A(\psi)$ for the right-hand side of (4.5) and solving for ψ. This yields

$$\hat{\psi} = \frac{ad}{bc},$$

the observed or empirical odds ratio, whose use in (4.11) leads to $A(\hat{\psi}) = a$ as required. It is reassuring that these somewhat abstract considerations have led to the obvious estimate in this simple case; in other more complicated situations the correct or "best" estimate is not at all obvious but may nevertheless be deduced from

analogous considerations. The empirical odds ratio is also the *unconditional* maximum likelihood estimate based on the two parameter product binomial distribution mentioned earlier.

Tests of significance

Large sample hypothesis tests are obtained *via* normal approximations to the tail probabilities (4.6) and (4.7) of the discrete conditional distribution. Figure 4.1 illustrates this process schematically. The approximating continuous distribution is first chosen to have the same mean and variance as the discrete one. Probabilities for the continuous distribution are represented by areas under the smooth curve, and for the discrete distribution by the areas of the rectangles centred over each possible value. Thus the exact probability in the right tail associated with the observed value 8 consists of the sum of the areas of the rectangles over 8, 9 and 10. It is clear from the diagram that this is best approximated by taking the area under the continuous curve from $7^1/_2$ to infinity. If we did not subtract $^1/_2$ from the observed value but instead took the area under the normal curve from 8 to infinity as an approximate p-value, we would underestimate the actual tail probability of the discrete distribution. More generally, if the values of the discrete distribution are spaced a constant distance Δ units apart, it would be appropriate *to reduce the observed value by* $^1/_2\,\Delta$ before referring it to a continuous distribution for approximation of an upper tail probability. Similarly, in the lower tail, the observed value would be increased by $^1/_2\,\Delta$. Such an adjustment of the test statistic is known as the *continuity correction.*

Since the hypergeometric distribution takes on integral values, for the problem at hand $\Delta = 1$. Thus the approximating tail probabilities may be written

$$p_U \approx 1-\Phi\left(\frac{a-A-{}^1/_2}{\sqrt{Var}}\right) \qquad (4.15)$$

and

$$p_L \approx \Phi\left(\frac{a-A+{}^1/_2}{\sqrt{Var}}\right)$$

where A and Var are the null mean and variance defined in (4.12) and (4.14), and Φ is the cumulative of the standard normal distribution. For a one-tailed test we generally report the upper tail probability, provided that the alternative hypothesis $\psi > 1$ has been specified before the study or it was the only one plausible. Similarly, for a one-tailed test against $\psi < 1$ we report p_L; however, for a two-tailed test, appropriate when the direction of the alternative cannot be specified in advance, we report twice the minimum value of p_L and p_U.

A convenient way of carrying out these calculations is in terms of the *corrected chi-square statistic*[1]*:*

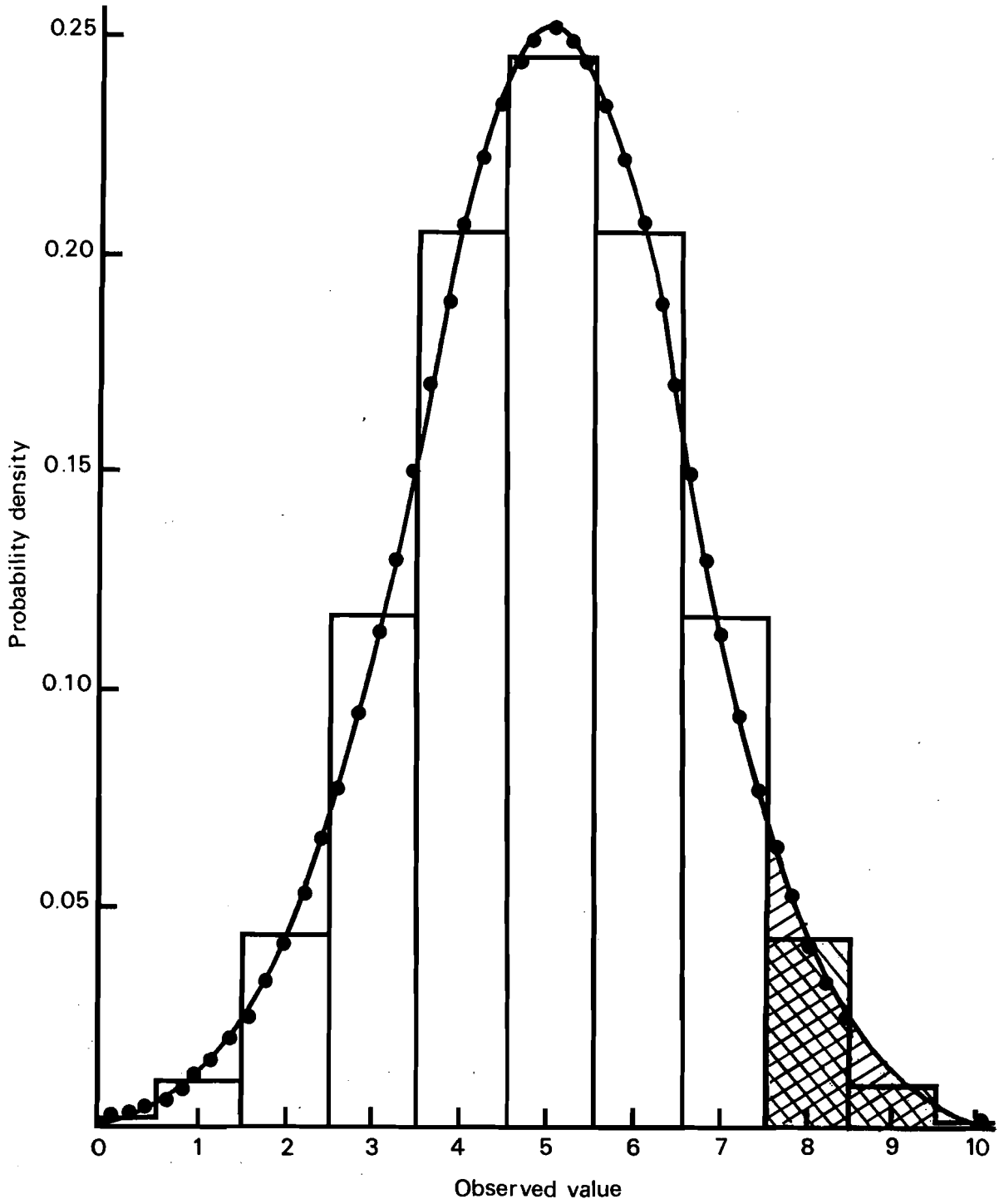
$$\chi^2 = \frac{(|a-A|-{}^1/_2)^2}{Var} = \frac{(|ad-bc|-{}^1/_2 N)^2\,(N-1)}{n_0 n_1 m_0 m_1}. \qquad (4.16)$$

[1] N-1 is often replaced by N in this expression.

Fig. 4.1 Normal approximation to discrete probability distribution. Note that the discrete probabilities for the values 8, 9 and 10 are better approximated by the area under the normal curve to the right of $7^1/_2$ than by the area under the normal curve to the right of 8.

▨ = normal distribution from $7^1/_2$ to infinity

▨ = discrete probabilities for the values 8, 9, and 10.



Referring this statistic to tables of percentiles of the chi-square distribution with one degree of freedom yields the approximate *two-sided* significance level, which may be halved to obtain the corresponding single-tail test.

There is no doubt that the $^1/_2$ continuity correction in (4.15) and (4.16) results in a closer approximation to the p-values obtained from the exact test discussed in the last section (Mantel & Greenhouse, 1968). Since the conditional distribution involves only the odds ratio as a parameter, and thus permits the derivation of point estimates, confidence intervals and significance tests in a unified manner, we feel it is the most appropriate one for assessing the evidence from any given set of data, and we therefore recommend the $^1/_2$ correction. This point, however, is somewhat controversial. Some authors argue that the exact test is inappropriate and show that more powerful tests can be constructed (for example, Liddell, 1978). These tests are not based on the conditional distribution, and their significance values are influenced by nuisance parameters.

It is important to recognize that when the sample is small we cannot rely on the asymptotic normal distribution to provide a reasonable approximation to the exact test. A general "rule of thumb" is that the approximations to significance levels in the neighbourhood of 0.05 or larger are reasonably good, providing that the *expected* frequencies for all four cells in the 2×2 tables are at least 5 under the null hypothesis (Armitage, 1971). These expectations may need to be considerably larger if p-values less than 0.05 are to be well approximated. For smaller samples, or when in doubt, recourse should be made to Fisher's exact test.

Cornfield's limits

Cornfield (1956) suggested that confidence intervals for the relative risk be obtained by approximating the discrete probabilities in (4.8) and (4.9). This leads to the equations

$$a - A(\psi_L) - {}^1/_2 = Z_{\alpha/2}\sqrt{\mathrm{Var}(a; \psi_L)}$$

and

$$a - A(\psi_U) + {}^1/_2 = -Z_{\alpha/2}\sqrt{\mathrm{Var}(a; \psi_U)} \qquad (4.17)$$

for the lower and upper limit, respectively. Here $Z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution (e.g., $Z_{.025} = 1.96$), while $A(\psi)$ and $\mathrm{Var}(a; \psi)$ are defined by (4.11) and (4.13). Cornfield's limits provide the best approximation to the exact limits (4.8) and (4.9), and come closest to achieving the nominal specifications (e.g., 95% confidence) of any of the confidence limits considered in this section (Gart & Thomas, 1972). Unfortunately the equations (4.17) are quartic equations which must be solved using iterative numerical methods. While tedious to obtain by hand, their solution has been programmed for a high-speed computer (Thomas, 1971).

The rule of thumb used to judge the adequacy of the normal approximation to the exact test may be extended for use with these approximate confidence intervals (Mantel & Fleiss, 1980). For 95% confidence limits, one simply establishes that the ranges $A(\psi_L) \pm 2\sqrt{\mathrm{Var}(a; \psi_L)}$ and $A(\psi_U) \pm 2\sqrt{\mathrm{Var}(a; \psi_U)}$ are both contained within the range of possible values for a. More accurate confidence limits are of course ob-

tained if the mean and variance of the exact conditional distribution are substituted in (4.17) for A and Var; however, this requires solution of polynomial equations of an even higher order.

Logit confidence limits

A more easily calculated set of confidence limits may be derived from the normal approximation to the distribution of $\log \hat{\psi}$ (Woolf, 1955). This has mean $\log \psi$ and a large sample variance which may be estimated by the sum of the reciprocals of the cell entries

$$\text{Var}(\log \hat{\psi}) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}. \tag{4.18}$$

Consequently, approximate $100(1-\alpha)\%$ confidence limits for $\log \psi$ are

$$\log \psi_U, \log \psi_L = \log \hat{\psi} \pm Z_{\alpha/2}\sqrt{\text{Var}(\log \hat{\psi})} \tag{4.19}$$

which may be exponentiated to yield ψ_L and ψ_U. Gart and Thomas (1972) find that such limits are generally too narrow, especially when calculated from small samples. Since $\log \hat{\psi}$ is the difference between two logit transformations (see Chapter 5), the limits obtained in this fashion are known as *logit limits.*

Test-based confidence limits

Miettinen (1976) has provided an even simpler and rather ingenious method for constructing confidence limits using only the point estimate and χ^2 test statistic. Instead of using (4.18), he solves

$$\frac{\log^2(\hat{\psi})}{\text{Var}(\log \hat{\psi})} = \chi^2,$$

for the variance of $\log(\hat{\psi})$, arguing that both left and right side provide roughly equivalent statistics for testing the null hypothesis $\psi = 1$. This technique is of even greater value in complex situations where significance tests may be fairly simple to calculate but precise estimates for the variance require more effort.

Substituting the test-based variance estimate into (4.19) yields the approximate limits

$$\psi_L, \psi_U = \hat{\psi}^{(1 \pm Z_{\alpha/2}/\chi)}, \tag{4.20}$$

where $\hat{\psi}$ is raised to the power $(1 \pm Z_{\alpha/2}/\chi)$. Whether ψ_L corresponds to the – sign in this expression and ψ_U to the + sign, or vice versa, will depend on the relative magnitude of $Z_{\alpha/2}$ and χ. The χ^2 statistic (4.16), however, should be calculated *without the continuity correction* especially when $\hat{\psi}$ is close to unity, since otherwise the variance may be overestimated and the limits too wide. In those rare cases where $\hat{\psi}$ is exactly equal to unity, the uncorrected χ^2 is equal to zero and the test-based limits are consequently undefined.

Halperin (1977) pointed out that the test-based variance estimate is strictly valid only if $\psi = 1$. When case and control sample sizes are equal ($n_1 = n_0$) the variance for other values of ψ is systematically underestimated by this approach, the true average probability is less than the nominal $100(1-\alpha)\%$, and the resulting confidence limits are too narrow (Gart, 1979). If there are large differences in the numbers of cases and controls, the true variance of $\log \hat{\psi}$ may sometimes be overestimated and the resulting limits will be too wide. Nevertheless the test-based limits may be advocated on the basis of their simplicity since they yield numerical results which are often in reasonable agreement with the other procedures and of sufficient accuracy for many practical purposes, at least when the estimated relative risk is not too extreme. They also provide convenient initial values from which to start the iterative solution of the equations for the more accurate limits, if these are desired.

Example: We illustrate these calculations with data from the 2×2 table shown in § 4.1. The (unconditional) maximum likelihood estimate of the relative risk is:

$$\hat{\psi} = \frac{96 \times 666}{104 \times 109} = 5.64,$$

while the corrected χ^2 test statistic is

$$\chi^2 = \frac{(|196 \times 666 - 104 \times 109| - {}^1/_2\, 975)^2 974}{200 \times 775 \times 770 \times 205} = 108.11$$

corresponding to a two-sided significance level of $p<0.0001$. The uncorrected χ^2 is slightly larger at 110.14. We use this latter value for determining the test-based 95% confidence intervals. These are

$$\psi_U, \psi_L = 5.64^{\,(1 \pm 1.96/\sqrt{110.14})} = 4.08, 7.79.$$

To calculate the logit limits we need

$$\mathrm{Var}(\log \hat{\psi}) = \frac{1}{96} + \frac{1}{104} + \frac{1}{109} + \frac{1}{666} = 0.0307,$$

leading to limits for $\log \psi$ of $\log 5.65 \pm 1.96 \times \sqrt{0.0307} = 1.730 \pm 0.343$, i.e., $\psi_L = 4.00$ and $\psi_U = 7.95$. By way of contrast, the Cornfield limits (4.17) yield $\psi_L = 3.94$ and $\psi_U = 8.07$.

For these data the logit limits are wider than the test-based limits, reflecting the fact that the estimated odds ratio is far from unity and the test-based variance is therefore too small. Both the logit and the test-based limits are too narrow in comparison with Cornfield's limits, but the magnitude of the discrepancy is not terribly great from a practical viewpoint. To gauge the accuracy of the normal approximation used to derive the Cornfield limits, following the procedure suggested by Mantel and Fleiss, we need to calculate the means and variances of the number of exposed cases under each of the two limits. The means are obtained as the solution for $A = A(\psi)$ in the quadratic equations (4.11)

$$\frac{A(775-205+A)}{(200-A)(205-A)} = \psi$$

for $\psi = \psi_L$ and ψ_U, namely:

$$A(\psi_L) = A(3.94) = 84.22$$

corresponding to fitted frequencies of

84.22	115.78	200
120.78	654.22	775
205	770	

and variance of

$$\text{Var}(a;\psi = 3.94) = \left[\frac{1}{84.22} + \frac{1}{115.78} + \frac{1}{120.78} + \frac{1}{654.22}\right]^{-1} = 32.98;$$

and

$$A(\psi_U) = A(8.07) = 107.48$$

with fitted frequencies

107.48	92.52	200
97.52	677.48	775
205	770	

and variance of

$$\text{Var}(a;\psi = 8.07) = \left[\frac{1}{107.48} + \frac{1}{92.52} + \frac{1}{97.52} + \frac{1}{677.48}\right]^{-1} = 31.40.$$

It is instructive to verify that the empirical odds ratios calculated from the fitted frequencies satisfy

$$\frac{84.22 \times 654.22}{115.78 \times 120.78} = 3.94 = \psi_L$$

and

$$\frac{107.48 \times 677.48}{97.52 \times 92.52} = 8.07 = \psi_U,$$

respectively. The actual range of possible values for a is max(0,205–775) to min(200,205), i.e., (0,200). This is much broader than the intervals including two standard deviations on both sides of the fitted means $84.22 \pm 2\sqrt{32.98} = (72.7, 95.7)$ and $107.48 \pm 2\sqrt{31.40} = (96.3, 118.7)$. Hence there is little doubt about the accuracy of the normal approximation for these data.

4.4 Combination of results from a series of 2 × 2 tables; control of confounding

The previous two sections dealt with a special situation which rarely occurs in practice. We have devoted so much attention to it in order to introduce, in a simplified setting, the basic concepts needed to solve more realistic problems, such as those posed by the presence of nuisance or confounding factors. Historically one of the most important methods for control of confounding has been to divide the sample into a series of strata which were internally homogeneous with respect to the confounding factors. Separate relative risks calculated *within* each stratum are free of bias arising from confounding (§ 3.4).

In such situations one first needs to know whether the association between exposure and disease is reasonably constant from stratum to stratum. If so, a summary measure of relative risk is required together with associated confidence intervals and tests of significance. If not, it is important to describe how the relative risk varies according to changes in the levels of the factors used for stratum formation. In this chapter we emphasize the calculation of summary measures of relative risk and tests of the hypothesis that it remains constant from stratum to stratum. Statistical models which

are particularly suited to evaluating and describing *variations* in relative risk are introduced in Chapters 6 and 7.

Example continued: Since incidence rates of most cancers rise sharply with age, this must always be considered as a potential confounding factor. We have already noted that the Ille-et-Vilaine cases were on average about ten years older than controls (Table 4.1). If age were also related to alcohol consumption, this would indicate that confounding existed in the data and we would expect to see the age-adjusted relative risk change accordingly. We know from Table 4.2, however, that age and alcohol are not strongly correlated, so that in this case the confounding effect may be minimal. Nevertheless we introduce age-stratification in order to illustrate the basic process.

Dividing the population into six 10-year age intervals yields the following series of 2×2 tables, whose sum is the single 2×2 table considered earlier (§ 4.1):

Age (years)		Daily alcohol consumption 80+ g	0–79 g	Odds ratio
25–34	Case	1	0	∞
	Control	9	106	
35–44	Case	4	5	5.05
	Control	26	164	
45–54	Case	25	21	5.67
	Control	29	138	
55–64	Case	42	34	6.36
	Control	27	139	
65–74	Case	19	36	2.58
	Control	18	88	
75+	Case	5	8	∞
	Control	0	31	

Some 0 cells occur in the youngest and oldest age groups, which have either a small number of cases or a small number of exposed. While these two tables do not by themselves provide much useful information about the relative risk, the data from them may nevertheless be combined with the data from other tables to obtain an overall estimate. There appears to be reasonable agreement between the estimated relative risks for the other four age groups, with the possible exception of that for the 65–74-year-olds.

A full analysis of such a series of 2×2 tables comprises: (1) a test of the null hypothesis that $\psi = 1$ in all tables; (2) point and interval estimation of ψ assumed to be common to all tables; and (3) a test of the homogeneity or no-interaction hypothesis

that ψ is constant across tables. Of course if this latter hypothesis is rejected, the results from (1) and (2) are of little interest. In this situation it is more important to try to understand and describe the sources of variation in the relative risk than simply to provide a summary measure.

The entries in the i^{th} of a series of I 2×2 tables may be identified as follows:

	Exposed	Unexposed	
Cases	a_i	b_i	n_{1i}
Controls	c_i	d_i	n_{0i}
	m_{1i}	m_{0i}	N_i

(4.21)

Conditional on fixed values for the marginal totals n_{1i}, n_{0i}, m_{0i} in each table, the probability distribution of the data consists of a product of I non-central hypergeometric terms of the form (4.2). A completely general formulation places no restriction on the odds ratios ψ_i in each table, but in most of the following we shall be working under the hypothesis that they are equal, $\psi_i = \psi$.

Summary chi-square: test of the null hypothesis

Under the hypothesis of no association, the expectation and variance of the number of exposed cases a_i in the i^{th} table are:

$$A_i(1) = \frac{n_{1i}m_{1i}}{N_i},$$

and (4.22)

$$\mathrm{Var}(a_i; \psi = 1) = \frac{n_{1i}n_{0i}m_{1i}m_{0i}}{N_i^2(N_i - 1)},$$

respectively (see equations 4.12 and 4.14). If the odds ratio is the same in each table, we would expect the a_i to be generally either larger ($\psi > 1$) or smaller ($\psi < 1$) than their mean values when $\psi = 1$. Hence an appropriate test is to compare the total $\Sigma_i a_i$ of the exposed cases with its expected value under the null hypothesis, dividing the difference by its standard deviation. The test statistic, corrected for the discrete nature of the data, may be written

$$\chi^2 = \frac{\left(\left| \sum_{i=1}^{I} a_i - \sum_{i=1}^{I} A_i(1) \right| - {}^1/_2 \right)^2}{\sum_{i=1}^{I} \mathrm{Var}(a_i; \psi = 1)}. \qquad (4.23)$$

This summary test was developed by Cochran (1954) and by Mantel and Haenszel (1959), with the latter suggesting use of the exact variances (4.22). Referring χ^2 to tables of the chi-square distribution with one degree of freedom provides two-sided

significance levels for evaluating the null hypothesis; these may be halved for a one-tail test.

Mantel and Fleiss (1980) suggest an extension of the "rule of 5" for evaluating the adequacy of the approximation to the exact p-value obtained from the summary chi-square statistic. They first calculate the maximum and minimum values that the total number of the exposed cases Σa_i may take subject to fixed marginals in each of the contributing 2×2 tables. These are $\Sigma \min(m_{1i}, n_{1i})$ for the maximum and $\Sigma \max(0, m_{1i} - n_{0i})$ for the minimum, respectively. Provided that the calculated mean value under the null hypothesis $\Sigma A_i(1)$ is at least five units away from both these extremes, the exact and approximate p-values should agree reasonably well for p's in the range of 0.05 and above. Similar considerations apply when evaluating the accuracy of the normal approximation in setting confidence limits (see equation 4.27 below). Here the mean values $\Sigma A_i(\psi)$ calculated at the confidence limits ψ_L and ψ_U for the odds ratio should both be at least $2\sqrt{\Sigma_i \mathrm{Var}(a_i; \psi)}$ units away from the minimum and maximum values.

Logit estimate of the common odds ratio[1]

Woolf (1955) proposed that a combined estimate of the log relative risk be calculated simply as a weighted average of the logarithms of the observed odds ratios in each table. The best weights are inversely proportional to the estimated variances shown in (4.18). Thus the logit estimate $\hat{\psi}_1$ is defined by

$$\log \hat{\psi}_1 = \frac{\Sigma w_i \log\left(\frac{a_i d_i}{b_i c_i}\right)}{\Sigma w_i} \tag{4.24}$$

$$\text{where } w_i = \left\{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right\}^{-1}.$$

The variance of such an estimate is given by the reciprocal of the sum of the weights, namely

$$\mathrm{Var}(\log \hat{\psi}_1) = \left(\Sigma w_i\right)^{-1}.$$

While the logit estimate behaves well in large samples, where all cell frequencies in all strata are of reasonable size, it runs into difficulty when the data are thin. For one thing, if any of the entries in a given table are 0, the log odds ratio and weight for that table are not even defined. The usual remedy for this problem is to add $^1/_2$ to each entry before calculating the individual odds ratios and weights (Gart & Zweifel, 1967; Cox, 1970). However the estimate calculated in this fashion is subject to unacceptable bias when combining information from large numbers of strata, each containing only a few cases or controls (Gart, 1970; McKinlay, 1978); thus it is not recommended for general use.

[1] This and the following subsections may be omitted at a first reading as they discuss, for the sake of completeness, estimates of the common odds ratio which are not used in the sequel.

Maximum likelihood estimate

The maximum likelihood estimate (MLE) of the common odds ratio is found by equating totals of the observed and expected numbers of exposed cases:

$$\sum_{i=1}^{I} a_i = \sum_{i=1}^{I} E(a_i \mid n_{1i}, n_{0i}, m_{1i}, m_{0i}; \psi). \tag{4.25}$$

For the exact or conditional MLE the expectations $E(a_i)$ are calculated under the non-central hypergeometric distributions (4.2), which means solution of a high degree polynomial equation. While the computational burden is thus sufficient to rule out use of this estimate for routine problems, a computer programme is available for special circumstances (Thomas, 1975). Calculation of the variance of the conditional MLE requires the variance of the conditional distribution and access to another computer programme (Zelen, 1971; Breslow, 1976).

For the unconditional MLE, based on the distribution of all the data without assuming fixed margins for each 2×2 table, the expectations $E(a_i)$ in (4.25) are those of the approximating normal distributions. Thus the estimation procedure requires finding fitted frequencies for all the cells, as in (4.10), such that the total of the observed and fitted numbers of exposed cases agree (Fienberg, 1977). While iterative calculations are also required here, they are generally less arduous than for the exact estimate and do not become any more complicated when the numbers in each cell are increased. As discussed in § 6.5, general purpose computer programmes for fitting logistic regression or log linear models may be used to find the unconditional MLE and estimates of its variance.

When there are many strata, each containing small numbers of cases and controls, the unconditional MLE is biased in the sense of giving values for ψ which are systematically more extreme (further from unity) than the true odds ratio. Numerical results on the magnitude of this bias in some special situations are given in Chapter 7. While the conditional MLE is not subject to this particular problem, it may be computationally burdensome even when there is ready access to an electronic computer. Hence none of the estimates considered so far are sufficiently simple or free of bias to be recommended for general use by the non-specialist.

The Mantel-Haenszel (M-H) estimate

Mantel and Haenszel (1959) proposed as a summary relative risk estimate the statistic

$$\hat{\psi}_{mh} = \frac{\sum_{i=1}^{I} a_i d_i / N_i}{\sum_{i=1}^{I} b_i c_i / N_i}, \tag{4.26}$$

which can be recognized as a weighted average of the individual odds ratios $\hat{\psi}_i = (a_i d_i)/(b_i c_i)$, with weights $b_i c_i / N_i$ which approximate the inverse variances of the individual estimates when ψ is near 1.

The Mantel-Haenszel (M–H) formula is not affected by zero cell entries and will give a consistent estimate of the common odds ratio even with large numbers of small

strata. When the data in each stratum are more extensive it yields results which are in good agreement with the MLEs (Gart, 1971; McKinlay, 1978). In view of its computational simplicity, it thus appears to be the ideal choice for the statistician or epidemiologist working with a pocket calculator on tabulated data. Its only major drawback is the lack of a robust variance estimate to accompany it.

Approximate confidence intervals

Exact confidence limits for this problem are discussed by Gart (1971) and have been programmed by Thomas (1975). Since their calculation is quite involved, however, we limit our discussion to the three types of approximate limits considered for the single 2×2 table. Following the same line of reasoning used to derive the Cornfield limits (4.17), normal approximations to the exact upper and lower $100(1-\alpha)\%$ confidence bounds are obtained as the solution of

$$\frac{\sum_{i=1}^{I} a_i - \sum_{i=1}^{I} A_i(\psi_U) + {}^1/_2}{\sqrt{\sum_{i=1}^{I} \mathrm{Var}(a_i; \psi_U)}} = -Z_{\alpha/2}$$

and (4.27)

$$\frac{\sum_{i=1}^{I} a_i - \sum_{i=1}^{I} A_i(\psi_L) - {}^1/_2}{\sqrt{\sum_{i=1}^{I} \mathrm{Var}(a_i; \psi_L)}} = Z_{\alpha/2},$$

respectively. Since $A_i(\psi)$ and $\mathrm{Var}(a_i; \psi)$ are defined as in (4.11) and (4.13), the calculation requires iterative solution of a series of quadratic equations (Thomas, 1975). The approximation is improved by use of the exact means and variances in place of the asymptotic ones. Though this requires even more calculation, use of the exact (conditional) moments is especially important when the number of strata is large and the data thin.

The logit limits are more easily obtained with a pocket calculator. These are defined by

$$\log \psi_U, \log \psi_L = \log \hat{\psi}_1 \pm Z_{\alpha/2} \Big/ \sqrt{\sum_{i=1}^{I} w_i}, \tag{4.28}$$

where $\hat{\psi}_1$ is the logit estimate and the w_i are the associated weights (4.24). Problems can be anticipated with their use in those same situations where the logit estimate has difficulties, namely when stratification becomes so fine that individual cell frequencies are small.

Miettinen's test-based limits require only a point estimate and test statistic. We recommend use of the M-H estimate for this purpose, and also use of the *uncorrected* version of χ^2. Thus

$$\psi_U, \psi_L = \hat{\psi}_{mh}^{(1 \pm Z_{\alpha/2}/\chi)}. \tag{4.29}$$

For reasons discussed in the previous section, these test-based limits become less accurate when the estimated relative risk is far from unity. They should not, however, be subject to the same tendency to increasing bias with increasing stratification as is the case with the logit limits.

Test for homogeneity of the odds ratio

All the procedures discussed so far in this section have been developed under the hypothesis that the odds ratio is constant across strata. If this were not the case, and a particular stratum had an odds ratio which was much larger than average, then we would expect the observed number of exposed cases a_i for that stratum to be larger than the expected number $A_i(\hat{\psi})$ based on the overall fitted odds ratio. Similarly, if the stratum odds ratio were small, we would expect a_i to be smaller than $A_i(\hat{\psi})$. Thus a reasonable test for the adequacy of the assumption of a common odds ratio is to sum up the squared deviations of observed and fitted values, each standardized by its variance:

$$\sum_{i=1}^{I} \frac{\{a_i - A_i(\hat{\psi})\}^2}{\mathrm{Var}(a_i; \hat{\psi})}. \tag{4.30}$$

If the homogeneity assumption is valid, and the size of the sample is large relative to the number of strata, this statistic follows an approximate chi-square distribution on I-1 degrees of freedom. While this is true regardless of which estimate $\hat{\psi}$ is inserted, use of the unconditional MLE has the advantage of making the total deviation $\Sigma_i\{a_i - A_i(\hat{\psi})\}$ zero. The statistic (4.30) is then a special case of the chi-square goodness of fit statistic for logistic models (§ 6.5); however, the M-H estimate also gives quite satisfactory results.

Unfortunately the global statistic (4.30) is not as useful as it may seem at first sight. If the number of strata is large and the data thinly spread out, the distribution of the statistic may not approximate the nominal chi-square even under the hypothesis of homogeneity. This is precisely the situation where the unconditional MLE breaks down. More importantly, even where it is valid the statistic may lack power against alternatives of interest. Suppose, for example, that the I strata correspond to values x_i of some continuous variable such as age and that the observed odds ratios systematically increase or decrease with advancing age. Such a pattern is completely ignored by the global test statistic, which is unaffected by the order of the strata. In such situations one should compute instead

$$\frac{\left[\sum_{i=1}^{I} x_i\{a_i - A_i(\hat{\psi})\}\right]^2}{\sum_{i=1}^{I} x_i^2 \mathrm{Var}\,(a_i; \hat{\psi}) - \left[\sum_{i=1}^{I} x_i \mathrm{Var}(a_i; \hat{\psi})\right]^2 \Big/ \sum_{i=1}^{I} \mathrm{Var}(a_i; \hat{\psi})}, \tag{4.31}$$

referring its value to tables of chi-square with one degree of freedom for a test of trend in ψ_i with x_i. If the x's are equally spaced, a continuity correction should be applied to the numerator of this statistic before squaring. Additional tests for trends in relative risk with one or several variables are easily carried out in the context of the modelling

approach. (In fact [4.31] is the "score" statistic for testing $\beta = 0$ in the model $\log \psi_i = \alpha + \beta x_i$. [See § 6.4, especially equation 6.18, and also § 6.12.])

In a similar context, suppose the I strata can be divided into H groups of size $I = I_1 + I_2 + \ldots + I_H$, and that we suspect the odds ratios are homogeneous within groups but not between them. Then, in place of the statistic (4.30) for testing overall homogeneity, we would be better advised to use

$$\sum_{h=1}^{H} \frac{\left[\sum_{i \epsilon I_h} a_i - A_i(\hat{\psi})\right]^2}{\sum_{i \epsilon I_h} \mathrm{Var}(a_i; \hat{\psi})}, \tag{4.32}$$

where the notation $\sum_{i \epsilon I_h}$ denotes summation over the strata in the h^{th} group. This statistic will be chi-square with only H-1 degrees of freedom under the homogeneity hypothesis, and has better power under the indicated alternative.

An alternative statistic for testing homogeneity, using the logit approach, is to take a weighted sum of the squared deviations between the separate estimates of log relative risk in each table and the overall logit estimate $\log \hat{\psi}_1$. This may be written

$$\sum_{i=1}^{I} w_i \log^2 \hat{\psi}_i - \left\{\sum_{i=1}^{I} w_i \log \hat{\psi}_i\right\}^2 \Big/ \sum_{i=1}^{I} w_i, \tag{4.33}$$

where the $\hat{\psi}_i$ denote the individual odds ratios and w_i the weights (4.24), both calculated after addition of $^1/_2$ to each cell entry. While this statistic should yield similar values to (4.30) when all the individual frequencies are large, it is even more subject to instability with thin data and is therefore not recommended for general practice.

Some other tests of homogeneity of the odds ratio which have been proposed are incorrect and should not be used (Halperin et al., 1977; Mantel, Brown & Byar, 1977). As an example we should mention the test obtained by adding the individual χ^2 statistics (4.16) for each table and subtracting the summary χ^2 statistic (4.23) (Zelen, 1971). This does not have an approximate chi-square distribution under the hypothesis of homogeneity unless all the odds ratios are equal to unity.

Example continued: Table 4.3 illustrates these calculations for the data for six age groups relating alcohol to oesophageal cancer introduced at the beginning of the section. The summary χ^2 statistic (4.23) for testing $\psi = 1$ is obtained from the totals in columns (2), (3) and (4) as

$$\chi^2 = \frac{(|96 - 48.890| - {}^1/_2)^2}{26.106} = 83.22,$$

which yields an equivalent normal deviate of $\chi = 9.122$, $p < 0.0001$. This is a slightly lower value than that obtained without stratification for age. Following the suggestion of Mantel and Fleiss (1980) for evaluating the adequacy of the normal approximation, we note that the minimum possible value for Σa_i consistent with the observed marginal totals is 0, while the maximum is 167, and that both extremes are sufficiently distant from the null mean of 48.890 to permit accurate approximation of p-values well below 0.05.

The logit estimate of the common odds ratio is calculated from the totals in columns (5) and (6) as

$$\log \hat{\psi}_1 = \frac{45.609}{28.261} = 1.614$$

i.e.,

$$\hat{\psi}_1 = \exp(1.614) = 5.022.$$

Similarly, from columns(7) and (8) we obtain the M-H estimate

$$\hat{\psi}_{mh} = \frac{58.439}{11.330} = 5.158.$$

By way of contrast, the conditional and asymptotic (unconditional) maximum likelihood estimates for this problem are $\hat{\psi}_{cond} = 5.251$ and $\hat{\psi}_{ml} = 5.312$, respectively. These numerical results confirm the tendency for the $^1/_2$ correction used with the logit estimate to result in some bias towards unity, and the opposite tendency for the estimate based on unconditional maximum likelihood. However, since the cell frequencies are of reasonably good size, except for the most extreme age groups, these tendencies are not large and all four estimates agree fairly well.

In practice one would report the estimated odds ratio with two or three significant digits, e.g., 5.2 or 5.16 for $\hat{\psi}_{mh}$. We have used more decimals here simply in order to illustrate the magnitude of the differences between the various estimates.

Ninety-five percent logit confidence limits, starting from the logit estimate, are

$$\log \psi_U, \psi_L = 1.614 \pm 1.96/\sqrt{28.261}$$

i.e.,

$$\psi_L = \exp(1.614 - 0.369) = 3.47$$
$$\psi_U = \exp(1.614 + 0.369) = 7.26.$$

However, since we know the logit point estimate is too small, these are perhaps best centered around $\log \hat{\psi}_{mh}$ instead, yielding

$$\psi_L = 5.158 \times \exp(-0.369) = 3.57$$
$$\psi_U = 5.158 \times \exp(0.369) = 7.46.$$

Test-based limits centred about the M-H estimate are computed from (4.29) as

$$\psi_L = 5.158^{(1-1.96/9.220)} = 3.64$$
$$\psi_U = 5.158^{(1+1.96/9.220)} = 7.31,$$

where $\chi = 9.220 = \sqrt{85.01}$ is the uncorrected test statistic, rather than the corrected value of 9.122. These limits are noticeably narrower than the logit limits. By way of contrast the Cornfield (4.27) limits $\psi_L = 3.60$ and $\psi_U = 7.84$, while yielding a broader interval (on the log scale) than either the logit or test-based limits, show the same tendency towards inflated values as does the unconditional maximum likelihood estimate. Thus, while all methods of calculation provide roughly the same value for the lower limit, namely 3.6, the upper limit varies between about 7.3 and 7.8.

In order to carry out the test for homogeneity we need first to find fitted values (4.11) for all the cell frequencies under the estimated common odds ratio. Using the (unconditional) MLE $\hat{\psi}_{ml} = 5.312$, we solve for A in the first table *via*

$$\frac{A(105 + A)}{(10-A)(1-A)} = 5.312,$$

which gives A(5.312) = 0.328. Fitted values for the remaining cells are calculated by subtraction, so as to yield a table with precisely the same marginal totals as the observed one, *viz:*

0.328	0.672	1
9.672	105.328	115
10	106	116

The variance estimate (4.13) is then

$$\text{Var}\,(a_i; \psi = 5.312) = \left(\frac{1}{0.328} + \frac{1}{0.672} + \frac{1}{9.672} + \frac{1}{105.328}\right)^{-1} = 0.215.$$

Table 4.3 Combination of data from a series of 2×2 tables

(1) Stratum (age in years)	(2) Data a	b	n_1	(3) (4) Test of null hypothesis[a]		(5) (6) Logit estimate[b]		(7) (8) Mantel-Haenszel estimate		(9) (10) Test of homogeneity[c]	
	c	d	n_0								
	m_1	m_2	N	A(1)	V(1)	$\log \hat{\psi}$	w	$\frac{ad}{N}$	$\frac{bc}{N}$	$A(\hat{\psi})$	$V(\hat{\psi})$
25–34	1	0	1	0.086	0.079	3.515	0.360	0.914	0.0	0.328	0.215
	9	106	115								
	10	106	116								
35–44	4	5	9	1.357	1.106	1.625	2.233	3.296	0.653	4.104	2.030
	26	164	190								
	30	169	199								
45–54	25	21	46	11.662	6.858	1.717	7.884	16.197	2.859	24.500	7.782
	29	138	167								
	54	159	213								
55–64	42	34	76	21.668	10.670	1.832	10.412	24.124	3.793	40.135	10.557
	27	139	166								
	69	173	242								
65–74	19	36	55	12.640	6.449	0.938	6.943	10.385	4.025	23.740	6.238
	18	88	106								
	37	124	161								
75+	5	8	13	1.477	0.944	3.708	0.429	3.523	0.0	3.203	0.996
	0	31	31								
	5	39	44								
Totals (except as noted)	96	104	200	48.890	26.106	45.609[d]	28.261	58.439	11.330	96.000	27.819
	109	666	775								
	205	770	975								

[a] Mean A(1) and variance V(1) of a under $\psi = 1$ from (4.12) and (4.14)
[b] Log relative risk estimates and weights from (4.24); $^1/_2$ added to each cell
[c] Mean and variance from (4.11) and (4.13) for $\hat{\psi} = \hat{\psi}_{ml}$, the unconditional MLE (4.25)
[d] Sum of $\log \hat{\psi}$ weighted by w

These values are listed at the head of columns (9) and (10) of Table 4.3; subsequent entries are calculated in precisely the same fashion for the other 2×2 tables. Thus the homogeneity chi-square (4.30) becomes

$$\frac{(1-0.328)^2}{0.215} + \frac{(4-4.104)^2}{2.030} + \ldots + \frac{(5-3.203)^2}{0.996} = 9.32$$

which when referred to tables of chi-square on I–1=5 degrees of freedom yields p = 0.10.

Notice that the total of the fitted values $A(\hat{\psi})$ in column (9) of Table 4.3 is precisely equal to the observed total, namely 96. In fact this relation is the defining characteristic of the unconditional MLE (see equation 4.25). If the fitted values are obtained instead from the M-H estimate, $\hat{\psi}_{mh} = 5.158$, they total 95.16 and give a value 9.28 to the homogeneity chi-square, very close to that already obtained. Thus it is perfectly feasible to carry out the test for homogeneity without having recourse to an *iteratively* computed estimate.

The alternative logit test statistic for homogeneity (4.33) is calculated as $0.360(3.515)^2 + 2.233(1.625)^2 + \ldots + 0.429(3.708)^2 - (45.609)^2/28.261 = 6.93$ (p=0.23). The reason this takes a smaller value is that for the extreme age categories, which contribute the most to the homogeneity chi-square, the addition of $^1/_2$ to the cell frequencies brings the odds ratios closer to the overall estimate.

Neither version of the formal test thus provides much evidence of heterogeneity. However, since the test lacks power it is important to continue the analysis by searching for patterns in the deviations between observed and fitted values which could indicate some sort of trend in the odds ratios. Certainly there is no obvious linear trend with age. This may be confirmed by assigning "dose" levels of $x_1 = 1$, $x_2 = 2$, ..., $x_6 = 6$ to the six age categories and computing the single degree of freedom chi-square for trend (4.31). We first need the intermediate quantities

$$\Sigma x_i\{a_i - A_i(\hat{\psi})\} = 1(1-0.328) + 2(4-4.104) + \ldots + 6(5-3.203) = -3.454$$

$$\Sigma x_i \mathrm{Var}\ (a_i;\hat{\psi}) = 1\times 0.215 + 2\times 2.030 + \ldots + 6\times 0.996 = 107.02$$

$$\Sigma x_i^2 \mathrm{Var}(a_i;\hat{\psi}) = 1\times 0.215 + 4\times 2.030 + \ldots + 36\times 0.996 = 439.09$$

from which we may calculate the test statistic

$$\frac{(-3.454 + {}^1/_2)^2}{439.09 - \dfrac{(107.02)^2}{27.819}} = 0.32.$$

When referred to tables of chi-square on one degree of freedom, this gives p = 0.58, i.e., no evidence for a trend.

4.5 Exposure at several levels: the $2 \times K$ table

The simple dichotomization of a risk variable in a 2×2 table with disease status will often obscure the full range of the association between exposure and risk. Qualitative exposure variables may occur naturally at several discrete levels, and more information can be obtained from quantitative variables if their values are grouped into four or five ordered levels rather than only two. Furthermore, this is the only way one can demonstrate the dose-response relationship which is so critical to the interpretation of an association as causal (§ 3.2). Hence there is a need for methods of analysis of several exposure levels analogous to those already considered for two levels.

Unstratified analysis

Suppose that there are $K > 2$ levels of exposure and that the subjects have been classified in a single $2 \times K$ table relating exposure to disease:

	Exposure level 1	2	...	K	Totals
Cases	a_1	a_2	...	a_K	n_1
Controls	c_1	c_2	...	c_K	n_0
Totals	m_1	m_2	...	m_K	N

(4.34)

The usual approach to data analysis in this situation is to choose one exposure level, say level 1, as a baseline against which to compare each of the other levels using the methods already given for 2×2 tables. In this way one obtains relative risks $r_1 = 1$, r_2, r_3, ..., r_K for each level, confidence intervals for these relative risks, and tests of the hypothesis that they are individually equal to unity.

To aid the interpretation of the results of such a series of individual tests, some of which may reach significance and others not, it is helpful to have available an overall test of the null hypothesis that the K relative risks r_k are all simultaneously equal to unity, i.e., that there is no effect of exposure on disease. Under this hypothesis, and conditional on the marginal totals n_1, n_0, m_1, ..., m_K, the numbers of cases exposed at the k^{th} level have the expectations

$$e_k = E(a_k) = \frac{m_k n_1}{N}, \tag{4.35}$$

variances

$$\mathrm{Var}(a_k) = \frac{m_k(N-m_k)n_1 n_0}{N^2(N-1)} \tag{4.36}$$

and covariances ($k \neq h$)

$$\mathrm{Cov}(a_k, a_h) = -\frac{m_k m_h n_1 n_0}{N^2(N-1)} \tag{4.37}$$

of the *K-dimensional hypergeometric distribution.* The test statistic itself is the usual one for testing the homogeneity of K proportions (Armitage, 1971), namely

$$\left(\frac{N-1}{N}\right) \sum_{k=1}^{K} (a_k - e_k)^2 \left\{\frac{1}{e_k} + \frac{1}{m_k - e_k}\right\} = (N-1)\left(\frac{1}{n_1} + \frac{1}{n_0}\right) \sum_{k=1}^{K} \frac{(a_k - e_k)^2}{m_k}, \tag{4.38}$$

which may be referred to tables of chi-square on K–1 degrees of freedom[1].

When the levels of the exposure variable have no natural order, as for a genetic polymorphism, this approach can be taken no further. However, for quantitative or ordered qualitative variables the overall chi-square wastes important information. A more sensitive way of detecting alternative hypotheses is to test for a *trend* in

[1] The leading term $\left(\frac{N-1}{N}\right)$ is often ignored.

disease risk with increasing levels of exposure. Suppose that there are "doses" x_k associated with the various levels of exposure, where we may simply take $x_k = k$ for an ordered variable. An appropriate statistic for testing trend is to consider the regression of the deviations $(a_k - e_k)$ on x_k (Armitage, 1955; Mantel, 1963). When squared and divided by its variance this becomes

$$\frac{N^2(N-1)\left\{\sum_{k=1}^{K} x_k(a_k - e_k)\right\}^2}{n_1 n_0 \left\{N \sum_{k=1}^{K} x_k^2 m_k - \left(\sum_{k=1}^{K} x_k m_k\right)^2\right\}}, \tag{4.39}$$

which should be referred to tables of chi-square on one degree of freedom. If the x_k are spaced one unit apart, as in the case of $x_k = k$, an appropriate correction for continuity is to reduce the *absolute value* of the numerator term $\Sigma x_k(a_k - e_k)$ by $^1/_2$ before squaring. Estimation of the quantitative trend parameter is best discussed in terms of the modelling approach of Chapter 6.

Adjustment by stratification

Confounding variables may be incorporated in the analysis of $2 \times K$ tables by stratification of the data just as described in § 4.4 for a 2×2 table. The frequencies of cases and controls classified by exposures for the i^{th} of I strata are simply expressed by the addition of subscripts to the entries in (4.34):

Stratum i

	Exposure level 1	2	...	K	Totals
Cases	a_{1i}	a_{2i}	...	a_{Ki}	n_{1i}
Controls	c_{1i}	c_{2i}	...	c_{Ki}	n_{0i}
Totals	m_{1i}	m_{2i}	...	m_{Ki}	N_i

(4.40)

Methods for analysis of a series of 2×2 tables may be used to estimate the adjusted relative risk for each level of exposure relative to the designated baseline level, to put confidence limits around this estimate, to test the significance of its departure from unity, and to test whether it varies from stratum to stratum. A peculiarity which results from this procedure when there is more than one $2 \times K$ table is that the estimated relative risks may not be consistent with each other. More precisely, if r_{21} is the summary estimate of the odds ratio comparing level 2 with level 1, and r_{31} the summary measure for level 3 compared with level 1, their ratio r_{31}/r_{21} is not algebraically identical to the summary odds ratio r_{32} comparing level 3 with level 2. This problem does not arise with a single table, since it is true for this case that

$$r_{31}/r_{21} = \frac{a_3c_1/a_1c_3}{a_2c_1/a_1c_2} = \frac{a_3c_2}{a_2c_3} = r_{32}\,.$$

Nor will it arise with a series of tables in which the relative risks comparing each pair of levels are the same from table to table. Therefore, inconsistency can be regarded as a particular manifestation of the problem of interaction (see § 5.5). Recourse must be made to the methods in Chapters 6 and 7 in order to have adjusted estimates of relative risk which display such consistency, and for general tests of interaction. Otherwise one is well advised to use as baseline category the one which contains the most information, i.e., the k such that the sum of the reciprocals of the numbers of cases and controls,

$$\sum_{i=1}^{I}\left\{\frac{1}{a_{ki}}+\frac{1}{c_{ki}}\right\},$$

is a minimum.

The generalization to stratified data of the statistic (4.38), which tests the global null hypothesis, is somewhat more complicated as it involves matrix manipulations (Mantel & Haenszel, 1959). Let us denote by $\mathbf{e}_i$ the K–1 dimensional vector of expectations $\mathbf{e}_i = E(\mathbf{a}_i) = E(a_{1i}, \ldots, a_{K-1,i})$ of numbers of cases exposed to each of the first K–1 levels in the i^{th} stratum, and by $\mathbf{V}_i$ the corresponding K–1 × K–1 dimensional covariance matrix. These are calculated as in formulae (4.35), (4.36), and (4.37) with the addition of i subscripts to all terms. Let $\mathbf{e}. = \Sigma\mathbf{e}_i$, $\mathbf{V}. = \Sigma\mathbf{V}_i$, and $\mathbf{a}. = \Sigma\mathbf{a}_i$ denote the sums of these quantities cumulated over the I strata. Then the global null hypothesis that there is no effect of exposure on disease, after adjustment by stratification, may be tested by referring the statistic

$$(\mathbf{a}.-\mathbf{e}.)^{T}\,\mathbf{V}.^{-1}(\mathbf{a}.-\mathbf{e}.) \tag{4.41}$$

to tables of chi-square with K–1 degrees of freedom. This reduces to (4.38) if I = 1, i.e., there is only a single stratum.

Calculation of this statistic requires matrix inversion, and while perfectly feasible to perform by hand for small values of K (say K=3 and 4), it becomes more difficult for larger values. Various approximate statistics have therefore been suggested (Armitage, 1966). One *conservative* approximation, which always yields values less than or equal to those of (4.41), is given by

$$\sum_{k=1}^{K}\frac{(a_{k.}-e_{k.})^2}{\sum_{i=1}^{I}\frac{n_{0i}e_{ki}}{N_i-1}}. \tag{4.42}$$

Unfortunately, the difference between (4.42) and (4.41) increases as the distributions of exposures among the combined case-control sample in each stratum become more disparate, which is one situation in which stratification may be important to reduce bias (Crowley & Breslow, 1975).

The statistic for the adjusted test of trend which generalizes (4.39) is more easily obtained as

$$\frac{\left\{\sum_{k=1}^{K}x_k(a_{k.}-e_{k.})\right\}^2}{\sum_{i=1}^{I}\frac{n_{0i}}{N_i-1}\left\{\sum_{k=1}^{K}x_k^2e_{ki}-\frac{1}{n_{1i}}\left(\sum_{k=1}^{K}x_ke_{ki}\right)^2\right\}}. \tag{4.43}$$

Here the numerator term represents the regression of the x's on the differences between the total observed and expected frequencies, while the denominator is its variance under the null hypothesis (Mantel, 1963). This statistic also should be referred to tables of chi-square with one degree of freedom, and a continuity correction applied to the numerator if the x_k values are equally spaced.

Example continued: As an illustration of the analysis of the effects of a risk factor taking on several levels, Table 4.4 presents data from Ille-et-Vilaine with alcohol consumption broken down into four levels rather than the two shown in § 4.1. Relative risks are calculated for each level of consumption against a baseline of 0–39 g/day as the empirical odds ratio for the corresponding 2×2 table. Each of these is individually highly significant as judged from the χ^2 test statistics, all of which exceed the critical value of 15.1 for significance at the $p = 0.0001$ level. Moreover, there is a clear increase in risk with increasing consumption. The confidence limits shown are those of Cornfield. It is perhaps worth remarking that the test-based limits are in better agreement with those for the lower levels (e.g., 2.29–5.57 for 40–79 g/day) than for the higher ones (16.22–45.92 for 120+ g/day), as would be expected in such a situation with a trend of risk.

While there is no doubt regarding the statistical significance of the observed differences in risk, and in particular the trend with increasing consumption, we nevertheless compute the chi-square test statistics (4.38) and (4.39) for purposes of illustration. The first step is the calculation of the table of expected values under the null hypothesis,

Table 4.4 Distribution of alcohol consumption for cases and controls: relative risks and confidence limits for each level, with and without adjustment for age

	Alcohol consumption (g/day)				Totals
	0–39	40–79	80–119	120+	
Cases	29	75	51	45	200
Controls	386	280	87	22	775
Totals	415	355	138	67	975
Unadjusted analysis					
RR ($\hat{\psi}$)	1.0	3.57	7.80	27.23	
χ^2	–	31.54	72.77	156.14	
95% confidence limit	–	2.21–5.77	4.54–13.46	13.8–54.18	
Global test of H_0: $\chi^2_3 = 158.8$			Test for trend: $\chi^2_1 = 151.9$		
Adjusted for age					
RR ($\hat{\psi}_{mh}$)	1.0	4.27	8.02	28.57	
RR ($\hat{\psi}_{ml}$)	1.0	4.26	8.02	37.82	
χ^2	–	36.00	57.15	135.49	
95% confidence limit	–	2.56–7.13	4.37–14.82	16.69–87.73	
Test for homogeneity (χ^3_5)		6.59	6.69	10.33	
Global test of H_0: $\chi^2_3 = 141.4$			Test for trend: $\chi^2_1 = 134.0$		

	Alcohol consumption (g/day) 0–39	40–79	80–119	120+	Totals
Cases	85.13	72.82	28.31	13.74	200.00
Controls	329.87	282.18	109.69	53.26	775.00
Totals	415.00	355.00	138.00	67.00	975.00

where the first row consists of the expected values e_k for cases and the second consists of the expected values $m_k - e_k$ for controls. Thus, for example,

$$e_1 = \frac{415 \times 200}{975} = 85.13$$

and $m_1 - e_1 = 415 - 85.13 = 329.87$. Note that the row and column totals of the observed and expected values agree. We then have from (4.38)

$$\frac{974}{975}\left\{(29-85.13)^2\left(\frac{1}{85.13}+\frac{1}{329.87}\right)+\ldots+(45-13.74)^2\left(\frac{1}{13.74}+\frac{1}{53.26}\right)\right\}=158.8,$$

which would normally be referred to tables of chi-square with three degrees of freedom.

Table 4.5 Distribution of alcohol consumption for cases and controls: in six age strata

Age (years)		Alcohol consumption (g/day) 0–39	40–79	80–119	120+	Total
25–34						
	Cases	0	0	0	1	1
	Controls	61	45	5	4	115
	Total	61	45	5	5	116
35–44						
	Cases	1	4	0	4	9
	Controls	88	76	20	6	190
	Total	89	80	20	10	199
45–54						
	Cases	1	20	12	13	46
	Controls	77	61	27	2	167
	Total	78	81	39	15	213
55–64						
	Cases	12	22	24	18	76
	Controls	77	62	19	8	166
	Total	89	84	43	26	242
65–74						
	Cases	11	25	13	6	55
	Controls	60	28	16	2	106
	Total	71	53	29	8	161
75+						
	Cases	4	4	2	3	13
	Controls	23	8	0	0	31
	Total	27	12	2	3	44

In calculating the chi-square for trend (4.39) we assign "doses" of $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, and $x_4 = 3$ to the four consumption levels, this assignment being justified on the grounds that the levels are more or less equally spaced. This yields

$$\frac{975^2 \times 974\ \{0(29-85.13)+\ldots+3(45-13.74)-{}^1/_2\}^2}{200 \times 775\ \{975(0\times 415+\ldots+9\times 67)-(0\times 415+\ldots+3\times 67)^2\}} = 151.9.$$

Hence most of the heterogeneity represented in the three degrees of freedom chi-square is "explained" by the linear increase in risk with dose.

In order to adjust these results for the possible confounding effects of age, we again stratify the population into six strata as shown in Table 4.5. Adjusted estimates of relative risk (Table 4.4) are obtained from the series of six 2×2 tables comparing each level with baseline, using the techniques already described in § 4.3. Since there was little correlation between age and alcohol consumption in the sample, and hence little confounding, the adjusted and unadjusted estimates do not differ much. If we calculate directly the relative risks for 120+ g/day *versus* 40–79 g/day using the series of six corresponding 2×2 tables we find a M-H summary odds ratio of $\hat{\psi}_{mh} = 8.71$ and MLE of $\hat{\psi}_{ml} = 9.63$. Neither of these agrees with the ratio of estimates for those two levels relative to 0–39 g/day shown in the table, i.e., 28.57/4.27 = 6.69 and 37.82/4.26 = 8.88, respectively. As mentioned earlier, the only way to achieve exact consistency among the summary measures is to build it into a modelling approach (Chapters 6 and 7).

The tendency of the unconditional MLE towards inflated values with thin data is evident for the 120 g/day category; in this case the conditional MLE is $\hat{\psi}_{cond} = 34.90$. Adjustment results in slightly less significant chi-squares and wider confidence limits, in accordance with the idea that "unnecessary" stratification leads to a slight loss of information or efficiency (§ 7.6). There is some evidence that the

Table 4.6 Expectations and covariances under the null hypothesis for the data in Table 4.5

Age (years)	Expected number of cases by level of alcohol (g/day) 0–39	40–79	80–119	120+	Covariance matrix[a] 0–39	40–79	80–119
					0.25	−0.20	−0.02
25–34	0.53	0.39	0.04	0.04	−0.20	0.24	−0.02
					−0.02	−0.02	0.04
					2.14	−1.55	−0.39
35–44	4.03	3.62	0.90	0.45	−1.55	2.08	−0.35
					−0.39	−0.35	0.78
					8.41	−5.05	−2.43
45–54	16.85	17.49	8.42	3.24	−5.05	8.54	−2.52
					−2.43	−2.52	5.42
					12.17	−6.68	−3.42
55–64	27.95	26.38	13.50	8.17	−6.68	11.86	−3.23
					−3.42	−3.23	7.65
					8.98	−5.29	−2.89
65–74	24.25	18.11	9.91	2.73	−5.29	8.05	−2.16
					−2.89	−2.16	5.38
					2.22	−1.57	−0.26
75+	7.98	3.55	0.59	0.89	−1.57	1.86	−0.12
					−0.26	−0.12	0.41
					34.17	−20.34	−9.41
Totals	81.59	69.54	33.36	15.52	−20.34	32.63	−8.40
					−9.41	−8.40	19.68

[a] The final row and column of this matrix, corresponding to the fourth level of 120+ g/day, are not shown as they are not needed for the subsequent calculations. They could be obtained from the fact that the sum of the matrix elements over any row or column is zero.

relative risk for the highest consumption level may vary with age, but the chi-square of 10.33 on five degrees of freedom does not quite attain nominal significance at p = 0.05, and considerable doubt exists as to the true significance level because of the small numbers in some tables. There is no evident trend in the relative risk with increasing age.

Expected values and covariances for the exposure frequencies of the cases *within* each stratum, calculated according to formulae (4.35)–(4.37), are presented in Table 4.6. For example, in the second stratum we have

$$\mathrm{Var}(a_{12}) = \frac{89 \times (199-89) \times 9 \times 190}{199 \times 199 \times 198} = 2.14$$

and

$$\mathrm{Cov}(a_{12},a_{22}) = -\frac{89 \times 80 \times 9 \times 190}{199 \times 199 \times 198} = -1.55.$$

The cumulated vector of expected exposures $\mathbf{e}_{\cdot}$ and covariance matrix $\mathbf{V}_{\cdot}$ are shown at the bottom of the table.

The adjusted global test (4.41) of the null hypothesis is calculated from the total observed values shown in Table 4.4 and the totals shown at the bottom of Table 4.6 as

$$(29\text{–}81.59,\ 75\text{–}69.54,\ 51\text{–}33.36) \begin{bmatrix} 0.108 & 0.091 & 0.091 \\ 0.091 & 0.111 & 0.091 \\ 0.091 & 0.091 & 0.133 \end{bmatrix} \begin{pmatrix} 29\text{–}81.59 \\ 75\text{–}69.54 \\ 51\text{–}33.36 \end{pmatrix} = 141.4,$$

where the 3×3 matrix is the inverse of the cumulated covariance matrix. To find the conservative approximation to this we compute from (4.42)

$$\frac{(29\text{–}81.59)^2}{\frac{1}{115}(0.53) + \frac{190}{198}(4.03) + \ldots + \frac{31}{43}(7.98)}$$

$$+ \quad \frac{(75\text{–}69.54)^2}{\frac{1}{115}(0.39) + \frac{190}{198}(3.62) + \ldots + \frac{31}{43}(3.55)}$$

$$\vdots$$

$$+ \quad \frac{(45\text{–}15.52)^2}{\frac{1}{115}(0.04) + \frac{190}{198}(0.45) + \ldots + \frac{31}{43}(0.89)}$$

$$= \quad 139.0.$$

In calculating the adjusted single degree of freedom test for trend (4.43), we first find the denominator terms

$$\Sigma x_k^2 e_{k1} = 0(0.53) + 1(0.39) + 4(0.04) + 9(0.04) = 0.91$$

$$\frac{(\Sigma\, x_k e_{k1})^2}{n_{11}} = \frac{\{0(0.53) + 1(0.39) + 2(0.04) + 3(0.04)\}^2}{1} = 0.35$$

$$\vdots \qquad\qquad \vdots$$

$$\Sigma x_k^2 e_{k6} = 0(7.98) + 1(3.55) + 4(0.59) + 9(0.89) = 13.92$$

$$\frac{(\Sigma\, x_k e_{k6})^2}{n_{16}} = \frac{\{0(7.98) + 1(3.55) + 2(0.59) + 3(0.89)\}^2}{13} = 4.21$$

and then use these in

$$\frac{\{0(29-81.59)+1(75-69.54)+\ldots+3(45-15.52)-{}^1/_2\}^2}{\frac{1}{115}(0.91-0.35)+\frac{190}{198}(11.27-5.09)+\ldots+\frac{31}{43}(13.92-4.21)}=134.0.$$

The test statistics are little affected by the adjustment process in this particular example, and the trend continues to account for the major portion of the variation[1].

Table 4.7 presents a summary of the results for tobacco analogous to those for alcohol shown in Table 4.4. While there is a clear association between an increased dose and increased risk, the relationship is not as strong as with alcohol nor does the linear trend component account for as much of it. In this case adjustment for age appears to increase the strength of the association, especially for the highest exposure category.

Table 4.7 Distribution of tobacco consumption for cases and controls: relative risks and confidence limits for each level, with and without adjustment for age

	Tobacco consumption (g/day)				
	0–9	10–19	20–29	30+	Total
Cases	78	58	33	31	200
Controls	447	178	99	51	775
Total	525	236	132	82	975
Unadjusted analysis					
RR ($\hat{\psi}$)	1.0	1.87	1.91	3.48	
χ^2		9.81	7.01	23.78	
95% confidence limit		1.25–2.78	1.17–3.11	2.03–5.96	
Global test of H_0: $\chi^2_3 = 29.3$			Test for trend: $\chi^2_1 = 26.9$		
Adjusted for age					
RR ($\hat{\psi}_{mh}$)	1.0	1.83	1.98	6.53	
RR ($\hat{\psi}_{ml}$)	1.0	1.85	1.99	6.60	
χ^2		8.29	6.76	37.09	
95% confidence limit		1.21–2.82	1.18–3.37	3.33–13.14	
Global test of H_0: $\chi^2_3 = 39.3$			Test for trend: $\chi^2_1 = 34.2$		

4.6 Joint effects of several risk factors

By defining each exposure category as a particular combination of factor levels, these same basic techniques can be used to explore the joint effects of two or more factors on disease risk. Relative risks are obtained using as baseline the category corresponding to the combination of baseline levels of each individual factor. Summary estimates of relative risk for one factor, adjusted for the effects of the others, are

[1] N.B. Since the intermediate results shown here are given only to two significant figures, whereas the exact values were used for calculation, some slight numerical discrepancies may be apparent when the reader works through these calculations himself.

obtained by including the latter among the stratification variables. Rather than attempt a discussion in general terms, details of this approach are best illustrated by a continuation of our analysis of the Ille-et-Vilaine data.

Example continued: The joint distribution of alcohol and tobacco consumption among cases and controls is shown in Table 4.8. Using the 0–9 g/day tobacco and 0–39 g/day alcohol categories as baseline, relative risks for each of the 15 remaining categories were obtained after stratification of the population into six age groups (Table 4.9). One of the difficulties of this method is that, as the data become more thinly spread out, an increasing fraction of the 2×2 tables from which the relative risks are calculated have at least one zero for a marginal total. This means that more and more data are effectively excluded from analysis since such tables make absolutely no contribution to any of the summary relative risk estimates or test statistics considered earlier. For example, only three out of the six tables contrasting the 30+ g/day tobacco, 120+ g/day alcohol exposure with the baseline level, namely those for the 45–54, 55–64 and 65–74 year age groups, were actually used to calculate the summary risk measure of 240.63. The remainder had at least one zero in a marginal total. This may explain the notable difference between the age-adjusted estimate and the crude relative risk estimate of $(10 \times 252)/(3 \times 9) = 93.33$. It is nevertheless apparent that people in this category of high alcohol/high tobacco consumption are at exceptional risk.

Table 4.9 shows a clear trend of increased risk with increased alcohol consumption within each tobacco category and likewise a trend with tobacco for each alcohol level. As neither of these variables accounts for the effects of the other, we say that they operate *independently* in producing their effects. Evidence for the lack of confounding in this instance comes from comparing the relative risks for alcohol which are simultaneously adjusted for age and tobacco (margin of Table 4.9) with those adjusted for age only Table 4.4). There is good agreement except perhaps for the highest level, where tobacco adjustment reduces the Mantel-Haenszel estimate from 28.6 to 22.8. Likewise the tobacco risks adjusted for alcohol and age do not depart greatly from those adjusted for age only (Table 4.7). Of course in other situations there may be risk factors which are partially confounded, some of their effect being due to the association with the other factor and some independent of it; and if there is complete confounding the effects of one may disappear after adjustment for the other.

Table 4.8 Joint classification of cases and controls by consumption of alcohol and tobacco

Alcohol (g/day)	Tobacco (g/day) 0–9		10–19		20–29		30+	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
0–39	9	252	10	74	5	35	5	23
40–79	34	145	17	68	15	47	9	20
80–119	19	42	19	30	6	10	7	5
120+	16	8	12	6	7	5	10	3

Table 4.9 Age-adjusted relative risks for joint exposure to alcohol and tobacco

Alcohol (g/day)	Tobacco (g/day) 0–9	10–19	20–29	30+	Adjusted for tobacco
0–39	1.0	3.90	4.17	9.44	1.0
40–79	8.18	8.63	10.57	52.47	4.05
80–119	12.94	13.88	17.97	155.62	7.49
120+	51.45	67.21	108.66	240.63	22.80
Adjusted for alcohol	1.0	1.51	1.56	8.10	

Computation of each of the simultaneously adjusted estimates shown in the margins of Table 4.9 involved the summarization of 24 2×2 tables, although many of these are omitted from the calculation because of zero marginals (for example, only 12 tables were used in the estimation of the relative risk of 8.10 for the highest tobacco level). Implicit in this calculation is the assumption that the odds ratios are constant over those tables being summarized, i.e., that the relative risks for tobacco do not depend on alcohol or age, while those for alcohol are independent of tobacco. Thus the relative risks shown in the margin of Table 4.9 are those obtained under the *multiplicative* hypothesis that the joint effect of alcohol and tobacco on incidence is the product of their individual effects (§ 2.6). *Smoothed* estimates of the relative risks for the combined categories under the multiplicative model are obtained by multiplying together the summary relative risks for each factor adjusting for the other. Thus the smoothed estimate for the 40–79 g/day alcohol, 10–19 g/day tobacco category is $1.51 \times 4.05 = 6.12$, compared with the individual cell estimate of 8.63.

Although we have shown that the method of stratification can be used to study the joint effects of two or more risk factors, it is not, in fact, well suited to this task. Computations become burdensome to perform by hand because so many strata must be created. Spreading the data out thinly may result in the loss of a large part of it from analysis. Hence, such multivariate analyses are best carried out using the regression models of Chapters 6 and 7, which permit a more economic, systematic and quantitative description of the effects of the several factors and their interactions.

REFERENCES

Armitage, P. (1955) Test for linear trend in proportions and frequencies. *Biometrics, 11,* 375–386

Armitage, P. (1966) The chi-square test for heterogeneity of proportions, after adjustment for stratification, *J. R. Stat. Soc. B., 28,* 150–163

Armitage, P. (1971) *Statistical Methods in Medical Research,* Oxford, Blackwell Scientific Publications

Breslow, N. (1976) Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics, 32,* 409–416

Cochran, W.G. (1954) Some methods for strengthening the common χ^2 tests. *Biometrics, 10,* 417–451

Cornfield, J. (1956) *A statistical problem arising from retrospective studies.* In: Neyman, J., ed., *Proceedings of the Third Berkeley Symposium, IV,* Berkeley, University of California Press, pp. 133–148

Cox, D.R. (1970) *The Analysis of Binary Data,* London, Methuen

Cox, D.R. & Hinkley, D.V. (1974) *Theoretical Statistics,* London, Chapman & Hall

Crowley, J. & Breslow, N. (1975) Remarks on the conservatism of $\Sigma(0\text{–}E)^2/E$ in survival data. *Biometrics, 31,* 957–961

Fienberg, S.E. (1977) *The Analysis of Cross-Classified Categorical Data,* Cambridge, Mass., MIT Press

Fleiss, J.L. (1973) *Statistical Methods for Rates and Proportions,* New York, Wiley

Gart, J.J. (1970) Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrics, 26,* 409–416

Gart, J.J. (1971) The comparison of proportions: a review of significance tests, confidence intervals, and adjustments for stratification. *Rev. Int. Stat. Inst., 39,* 148–169

Gart, J.J. (1979) Statistical analyses of the relative risk. *Environ. Health Perspect., 32,* 157–167

Gart, J.J. & Thomas, D.G. (1972) Numerical results on approximate confidence limits for the odds ratio, *J. R. Stat. Soc. B., 34,* 441–447

Gart, J.J. & Zweifel, J.R. (1967) On the bias of the logit and its variance, with application to quantal bioassay. *Biometrika, 54,* 181–187

Halperin, M. (1977) Letter to the Editor. *Am. J. Epidemiol., 105,* 496–498

Halperin, M., Ware, J.H., Byar, D.P., Mantel, N., Brown, C.C., Koziol, J., Gail, M. & Green, S.B. (1977) Testing for interaction in an $I \times J \times K$ contingency table. *Biometrika, 64,* 271–275

Hannan, J. & Harkness, W.L. (1963) Normal approximation to the distribution of two independent binomials, conditional on fixed sum. *Ann. Math. Stat., 34,* 1593–1595

Liddell, D. (1978) Practical tests of the 2×2 contingency tables. *Statistician, 25,* 295–304

Mantel, N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc., 58,* 690–700

Mantel, N. & Fleiss, J.L. (1980) Minimum requirements for the Mantel-Haenszel one-degree of freedom chi-square test and a related rapid procedure. *Am. J. Epidemiol.* (in press)

Mantel, N. & Greenhouse, S.W. (1968) What is the continuity correction? *Am. Stat. 22,* 27–30

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J. natl Cancer Inst., 22,* 719–748

Mantel, N., Brown, C. & Byar, D.P. (1977) Tests for homogeneity of effect in an epidemiologic investigation. *Am. J. Epidemiol., 106,* 125–129

McKinlay, S.M. (1978) The effect of non-zero second-order interaction on combined estimators of the odds ratio. *Biometrika, 65,* 191–202

Miettinen, O.S. (1976) Estimability and estimation in case-referrent studies. *Am. J. Epidemiol., 103,* 226–235

Rothman, K. & Boice, J. (1979) *Epidemiologic analysis with a programmable calculator (NIH Publication No. 79–1649),* Washington DC, US Government Printing Office

Thomas, D.G. (1971) Exact confidence limits for an odds ratio in a 2×2 table. *Appl. Stat., 20,* 105–110

Thomas, D.G. (1975) Exact and asymptotic methods for the combination of 2×2 tables. *Comput. Biomed. Res., 8,* 423–446

Tuyns, A.J., Péquignot, G. & Jensen, O.M. (1977) Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. *Bull. Cancer, 64,* 45–60

Woolf, B. (1955) On estimating the relationship between blood group and disease. *Ann. Human Genet., 19,* 251–253

Zelen, M. (1971) The analysis of several 2×2 contigency tables. *Biometrika, 58,* 129–137

LIST OF SYMBOLS – CHAPTER 4 (in order of appearance)

ψ	odds ratio (approximate relative risk)
a	number of exposed cases
b	number of unexposed cases
c	number of exposed controls
d	number of unexposed controls
n_1	number of cases (subtotal)
n_0	number of controls (subtotal)
m_1	number of exposed (subtotal)
m_0	number of unexposed (subtotal)
N	total number of cases and controls
P_1	probability of disease development for exposed
P_0	probability of disease development for unexposed
p_1	probability of exposure for a case
p_0	probability of exposure for a control
H_0	the null hypothesis that exposure has no effect on risk ($\psi = 1$)
$\binom{n}{u}$	binomial coefficient (see p. 125); there are $\binom{n}{u}$ ways of choosing u objects from n objects
pr ( )	the probability of an event ( )
pr (|)	the probability of one event conditional on another
$\hat{\psi}_{cond}$	conditional maximum likelihood estimate of the common odds ratio
E (|)	expectation of one random variable conditional on the values of another
H	a statistical hypothesis regarding the value of some parameter, for example $\psi = \psi_0$
p_L	lower tail probability or p-value
p_U	upper tail probability or p-value
ψ_L	lower confidence limit on the odds ratio
ψ_U	upper confidence limit on the odds ratio
α	size of a statistical test, predetermined significance level such that if the p-value falls below α one rejects the hypothesis
$A = A(\psi)$	expected number (asymptotic) of exposed cases when marginal totals of the 2×2 table are fixed, when the true odds ratio is ψ; fitted value for number of exposed cases when the true odds ratio is ψ
B, C, D	fitted values for remaining entries in the 2×2 table
$Var = Var(a;\psi)$	variance (asymptotic) of the number a of exposed cases when the marginal totals of the 2×2 table are fixed and the true odds ratio is ψ
$\hat{\psi}$	an estimate of the odds ratio
Δ	distance between adjacent observations of a discrete distribution (assumed constant)
Φ	cumulative distribution function of the standard normal distribution, e.g., $\Phi(-1.96) = 0.025$, $\Phi(1.96) = 0.975$
$\|x\|$	absolute value of a number x; the positive part of x; $\|3\| = \|-3\| = 3$

$Z_{\alpha/2}$	the $100(1-\alpha/2)$ percentile of the standard normal distribution: $\Phi(Z_{\alpha/2}) = 1-\alpha/2$
log	the natural logarithm; log to the base e
χ^2	a statistic which has (asymptotically) a chi-square distribution under the null hypothesis
χ	the square root of a χ^2 statistic
i	subscript added to denote the i^{th} stratum, e.g., a_i = number of exposed cases in the i^{th} stratum, ψ_i odds ratio in i^{th} stratum, etc.
$\hat{\psi}_l$	"logit" estimate of the common odds ratio in a series of 2×2 tables
w_i	weights associated with the logit estimate in the i^{th} stratum
$\hat{\psi}_{ml}$	(unconditional) maximum likelihood estimate (MLE) of the odds ratio
$\hat{\psi}_{mh}$	Mantel-Haenszel (M-H) estimate of the common odds ratio in a series of 2×2 tables
a_k	number of cases exposed to level k of a polytomous factor
c_k	number of controls exposed to level k of a polytomous factor
m_k	number of subjects (cases + controls) exposed to level k of a polytomous factor
e_k	expected number of cases exposed to level k under the null hypothesis and assuming fixed marginals in a $2 \times K$ table
Cov(x,y)	covariance between two variables x and y
Var(x)	variance of a variable x
$\mathbf{e}_i$	vector of expected values of the numbers of cases exposed to the first K–1 levels of a polytomous factor in the i^{th} stratum
$\mathbf{V}_i$	variance-covariance matrix of the numbers of cases exposed to the first K–1 levels of a polytomous risk factor in the i^{th} stratum
.	denotes summation over the subscript which it replaces; e.g., for the doubly subscripted array $\{a_{ki}\}$, $a_{k.} = \Sigma_i a_{ki} = a_{k1} + \ldots + a_{kI}$
χ^2_ν	a statistic which has (asymptotically) a chi-square distribution with ν degrees of freedom under the null hypothesis