5. CLASSICAL METHODS OF ANALYSIS OF MATCHED DATA

CHAPTER V

CLASSICAL METHODS OF ANALYSIS OF MATCHED DATA

As a technique for the control of confounding, stratification may be introduced either at the design stage of a study or during the analysis of results. An advantage of using it in design, keeping a constant ratio of controls to cases in each stratum, is that one avoids the inefficiencies resulting from having some strata with a gross imbalance of cases and controls. In the Ille-et-Vilaine study, for example, the 115 controls ascertained between 25 and 34 years of age are effectively lost from the analysis, or make only a minimal contribution to it, because there is only a single case with which to compare them (Table 4.1). Of course such gains in efficiency are only achieved if the analysis takes proper account of the stratification, which must be done in general anyway in order to avoid biased estimates of the relative risk (§ 3.4).

The ultimate form of a stratified design occurs when each case is individually matched to a set of controls, usually one or two but sometimes more, chosen to have similar values for certain of the important confounding variables. Some choices of control population intrinsically imply a matched design and analysis, as with neighbourhood or familial controls. If the exposure levels of the risk factor to be analysed are dichotomous or polytomous, the tests and estimates developed in the last chapter may be employed directly by considering each matched pair or set to be a separate stratum. Of course those "asymptotic" techniques which lead to trouble with sparse data should be avoided, while some of the "exact" procedures which were not considered feasible with general strata are quite tractable and useful with matched data. In this chapter we take advantage of the special structure imposed by the matching, so as to express many of the previously discussed tests and estimates in simple and succinct form.

5.1 Los Angeles retirement community study of endometrial cancer

An example which we shall use to illustrate the methods for matched data analysis is the study of the effect of exogenous oestrogens on the risk of endometrial cancer reported by Mack et al. (1976). These investigators identified 63 cases of endometrial cancer occurring in a retirement community near Los Angeles, California (USA) from 1971 to 1975. Each case was matched to four control women who were alive and living in the community at the time the case was diagnosed, who were born within one year of the case, who had the same marital status and who had entered the community at approximately the same time. In addition, controls were chosen from among women who had not had a hysterectomy prior to the time the case was diagnosed, and who were therefore still at risk for the disease.

Information on the history of use of several specific types of medicines, including oestrogens, anti-hypertensives, sedatives and tranquilizers, was abstracted from the medical record of each case and control. Other abstracted data relate to pregnancy history, mention of certain diseases, and obesity. Table 5.1 summarizes the distribution of cases and controls according to some of the key variables. Note the almost perfect balance of the age distribution of cases and controls, a consequence of the matching.

The analysis of these data is aimed at studying the risk associated with the use of oestrogens as well as with a history of gall bladder disease, and how these risks may be modified by the other factors shown in Table 5.1. When illustrating methods which involve matching a single control to a single case, the first of the four selected controls is used. A listing of the complete set of data is presented in Appendix III.

Table 5.1 Characteristics of cases and controls in Los Angeles study of endometrial cancer

Variable	Level	Cases	Controls	RR[a]
Age (years)	55–59	1	4	
	60–64	12	43	
	65–69	15	60	
	70–74	21	77	
	75–79	6	37	
	80+	8	31	
Mean		70.7	70.8	
S.D.		6.4	6.2	
Gall-bladder disease	Yes	17	24	3.5
	No	46	228	1.0
Hypertension	Yes	26	82	1.5
	No	37	170	1.0
Obesity	Yes	41	126	1.6
	No	16	81	1.0
	Unk	6	45	0.7
Other drugs (non-oestrogen)	Yes	56	176	3.5
	No	7	76	1.0
Oestrogens (any)	Yes	56	127	7.9
	No	7	125	1.0
Conjugated oestrogen: amount (mg/day)	None	12	143	1.0
	0.1–0.299	16	45	4.2
	0.3–0.625	15	41	4.4
	0.626+	16	19	10.0
	Unk	4	4	11.9
Conjugated oestrogen: duration (months)	None	12	143	1.0
	1–11	6	26	2.8
	12–47	12	32	4.5
	48–95	10	17	7.0
	96+	17	23	8.8
	Unk	6	11	6.5

[a] Relative risks calculated from unmatched data; RR = 1.0 identifies baseline category

5.2 Matched pairs: dichotomous exposure

The simplest example of matched data occurs when there is 1:1 pair matching of cases with controls and a single binary exposure. This is a special case of the situation considered in § 4.4, wherein each stratum consists of one case-control pair. The possible outcomes are represented by four 2×2 tables:

	Exposure									
	+	−	+	−	+	−	+	−	Total	
Case	1	0	1	0	0	1	0	1	1	(5.1)
Control	1	0	0	1	1	0	0	1	1	
Total	2	0	1	1	1	1	0	2	2	
Number of such tables:	n_{11}		n_{10}		n_{01}		n_{00}			

The most suitable statistical model for making inferences about the odds ratio with matched or very finely stratified data is to determine the conditional probability of the number of exposed cases in each stratum, assuming that the marginal totals of that stratum are fixed (§ 4.2). For tables in which there are zero marginal totals, i.e., for the extreme tables in which either both or neither the case or control are exposed to the risk factor, this conditional distribution assigns a probability of one to the observed outcome and hence contributes no information about the odds ratio. The statistical analysis uses just the *discordant* pairs, in which only the case or only the control is exposed. Denoting by $p_1 = 1\text{-}q_1$ and $p_0 = 1\text{-}q_0$ the exposure probabilities for case and control, respectively, the probability of observing a case-control pair with the case only exposed is p_1q_0 while that of observing a pair where only the control is exposed is q_1p_0. Hence the conditional probability of observing a pair of the former variety, given that it is discordant, is

$$\pi = \frac{p_1q_0}{p_1q_0 + q_1p_0} = \frac{\psi}{\psi + 1}, \tag{5.2}$$

a function of the odds ratio ψ. This is a special case of the general formula (4.2) in which $a = n_1 = n_0 = m_1 = m_0 = 1$. It follows that the conditional probability of observing n_{10} pairs with the case exposed and control not, conditional on there being $n_{10} + n_{01}$ discordant pairs total, is given by the binomial formula with probability parameter π

$$\text{pr}(n_{10} \mid n_{10} + n_{01}; p_0, p_1) = \binom{n_{10} + n_{01}}{n_{10}} \pi^{n_{10}} (1-\pi)^{n_{01}}. \tag{5.3}$$

While we will derive all statistical procedures for making inferences about ψ directly from this distribution, many can also be viewed as specializations of the general methods developed in § 4.4 for stratified samples.

Test of the null hypothesis

When $\psi = 1$, i.e., there is no association, the probabilities of the two different kinds of discordance are equal. Hence for small samples, say either n_{10} or n_{01} smaller than ten, the null hypothesis H_0: $\psi = 1$ may be tested by calculating the exact tail probabilities of the binomial distribution with probability $\pi = {}^1/_2$. Otherwise we use the continuity corrected version of the chi-square statistic based on the standardized value of n_{10}:

$$\chi^2 = \frac{\{|n_{10}-E(n_{10})|-{}^1/_2\}^2}{\mathrm{Var}(n_{10})} = \frac{\{|n_{10}-\frac{n_{10}+n_{01}}{2}|-{}^1/_2\}^2}{{}^1/_4\,(n_{10}+n_{01})}, \tag{5.4}$$

which is a special case of (4.23). Known as McNemar's (1947) test for the equality of proportions in matched samples, it is often expressed

$$\chi^2 = (|n_{10}-n_{01}|-1)^2/(n_{10}+n_{01}). \tag{5.5}$$

Estimating the odds ratio

Since the maximum likelihood estimate (MLE) of the binomial parameter π is simply the observed proportion of discordant pairs in which the case is exposed, it follows that the MLE of ψ is

$$\hat{\psi} = \frac{n_{10}}{n_{01}}, \tag{5.6}$$

i.e., the ratio of the two types of discordant pairs. This is essentially the only instance when the conditional MLE (4.25) discussed for stratified data can readily be calculated. It is interesting that $\hat{\psi}$ is also the Mantel-Haenszel (M-H) estimate (4.26) applied to matched pair data.

Confidence limits

Exact $100(1-\alpha)\%$ confidence intervals for the binomial parameter π in (5.2) may be determined from the charts of Pearson and Hartley (1966). Alternatively, they may be computed from the tail probabilities of the binomial distribution, using the formulae

$$\pi_L = \frac{n_{10}}{n_{10}+(n_{01}+1)F_{\alpha/2}(2n_{01}+2,2n_{10})}$$

and

$$\pi_U = \frac{(n_{10}+1)F_{\alpha/2}(2n_{10}+2,2n_{01})}{n_{01}+(n_{10}+1)F_{\alpha/2}(2n_{10}+2,2n_{01})}. \tag{5.7}$$

Here $F_{\alpha/2}(\nu_1,\nu_2)$ denotes the upper $100(\alpha/2)$ percentile of the F distribution with ν_1 and ν_2 degrees of freedom, in terms of which the cumulative binomial distribution may be expressed (Pearson & Hartley, 1966).

Approximate confidence limits for π are based on the normal approximation to the binomial tail probabilities. These are computed from the quadratic equations

$$\frac{n_{10}-\pi_L(n_{10}+n_{01})-{}^1/_2}{\sqrt{(n_{10}+n_{01})\pi_L(1-\pi_L)}} = Z_{\alpha/2}$$

and (5.8)

$$\frac{n_{10}-\pi_U(n_{10}+n_{01})+{}^1/_2}{\sqrt{(n_{10}+n_{01})\pi_U(1-\pi_U)}} = -Z_{\alpha/2}$$

where $Z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the normal distribution.

Once limits for π are found, whether from (5.7) or (5.8), they are converted into limits for ψ by using the inverse transformation

$$\psi = \frac{\pi}{1-\pi}. \tag{5.9}$$

Alternatively, substituting for π_L and π_U in (5.8), one can write the equations somewhat more simply as

$$\frac{n_{10}-\psi_L n_{01}-{}^1/_2(1+\psi_L)}{\sqrt{(n_{10}+n_{01})\psi_L}} = Z_{\alpha/2}$$

(5.10)

$$\frac{n_{10}-\psi_U n_{01}+{}^1/_2(1+\psi_U)}{\sqrt{(n_{10}+n_{01})\psi_U}} = -Z_{\alpha/2}$$

and solve directly for ψ_L and ψ_U.

Adjustment for confounding variables

One problem which occurs frequently in practice is that of adjusting for the confounding effects of a variable on which cases and controls have not been matched. In a study of the effects of a particular occupational exposure on lung cancer, for example, cases and controls may be matched on age and calendar year of diagnosis but not on smoking history. It would have been standard procedure in the past to adjust for the smoking effects by restricting the analysis to those case-control sets which were homogeneous for smoking according to some prescribed definition. Depending upon the stringency of the criteria for "same smoking history", this procedure could well result in the loss of a major portion of the data from analysis and is therefore wasteful. A much more satisfactory technique for control of confounding in a matched analysis is to model the effects of the confounding variables in a multivariate equation which also includes the exposures of interest (see § 7.2).

Testing for heterogeneity of the relative risk

It is important to note that the modifying effect of a variable is not altered by its use for case-control matching. Interaction effects can be estimated just as well from

matched as from unmatched data. For example, if both the incidence of the disease and the prevalence of a confounding variable vary throughout the region of study, one might well choose controls matched for place of residence. It would be appropriate and prudent to investigate if the relative risk associated with the exposure of interest was the same throughout the region. Partitioning the matched case-control pairs into subgroups on the basis of the variable of interest, in this case place of residence, enables separate relative risk estimates to be calculated for each subgroup and compared.

This approach could also be used to study the interaction effects of variables besides those used for matching. But, it then entails the same loss of information noted to occur when controlling for the confounding effects of such variables, since the analysis must be restricted to matched sets which are homogeneous for the additional variable(s).

With 1:1 pair matching the easiest way to test for the homogeneity of the odds ratios ψ in several subgroups is in terms of the associated probabilities π defined by (5.2). With H separate subgroups, one simply arranges the frequencies of discordant pairs in a $2 \times H$ table

	Subgroup 1	2	...	H
n_{10}				
n_{01}				

and carries out the appropriate test for independence or trend (see § 4.5). More advanced and flexible techniques for modelling interaction effects are presented in Chapter 7.

Example: We begin the illustrative analysis of the Los Angeles endometrial cancer study by confining attention to the first of the four controls and considering exposure as "ever having taken any oestrogen". This yields the following distribution of the 63 case-control pairs:

		Control: Exposed	Control: Non-exposed
Case	Exposed	27	29
	Non-exposed	3	4

Hence the ML estimate of the relative risk is 29/3 = 9.67 and the statistic (5.4) for testing the null hypothesis is

$$\chi^2 = \frac{\{|29-3|-1\}^2}{32} = 19.53,$$

corresponding to a significance level of p = 0.000005.

Ninety-five percent confidence limits based on the exact binomial distribution (5.7) are.

$$\pi_L = \frac{29}{29+4(2.42)} = 0.75 \text{ corresponding to } \psi'_L = 3.0$$

and

$$\pi_U = \frac{30(4.96)}{3+30(4.96)} = 0.97 \text{ corresponding to } \psi'_U = 49.6$$

where 2.42 = $F_{.025}$ (8,58) and 4.96 = $F_{.025}$ (60,6). Limits based on the normal approximation are found as solutions to the equations (5.10)

$$29-3\psi'_L-{}^1/_2(1+\psi'_L) = 1.96\sqrt{32\psi'_L}$$

and

$$29-3\psi'_U+{}^1/_2(1+\psi'_U) = -1.96\sqrt{32\psi'_U}\ ,$$

the solutions being $\psi'_L = 2.8$ and $\psi'_U = 39.7$, respectively.

Similar calculations may be made for the effect of a history of gall-bladder disease on endometrial cancer incidence. Here the overall matched pair data are

Case \ Control	+	−
+	4	13
−	5	41

RR = 13/5 = 2.6
χ^2 = 2.72 (p = 0.10)

Dividing the pairs according to the age of the case (and hence also the control) we find

Under 70 years

Case \ Control	+	−
+	2	7
−	1	18

RR = 7.0
p = 0.07

70 years and older

Case \ Control	+	−
+	2	6
−	4	23

RR = 1.5
p' = 0.75

where the two p-values were obtained from the tail probabilities of the binomial distribution with $\pi = {}^1/_2$ in view of the small numbers. To test for the homogeneity of the two relative risks in the different age groups we form the 2 × 2 table

	Age <70	Age ≧70	
n_{10}	7	6	13
n_{01}	1	4	5
	8	10	18

for which the usual (corrected) chi-square is χ^2 = 0.59, p = 0.44. Thus there is no evidence for a modifying effect of age on the relative risk for gall-bladder disease.

If we try to evaluate hypertensive disease as a confounding or modifying factor in a similar fashion, we find there is a severe loss of data because of the restriction to case-control pairs which are homogeneous for hypertension:

Hypertensive positive

Case \ Control	+	−
+	1	1
−	0	6

Hypertensive negative

Case \ Control	+	−
+	2	6
−	1	15

Only 32 of the original 63 pairs are available to estimate the relative risk associated with gall bladder disease while controlling for hypertension, and the number of discordant pairs actually used in the estimation is reduced from 18 to 8. As a measure of relative risk adjusted for hypertension we thus calculate

$$RR = 7/1 = 7.0$$

and for an adjusted test of the null hypothesis

$$\chi^2 = \frac{\{|7-1|-1\}^2}{8} = 3.13.$$

There is clearly almost no information left about how hypertension may modify the effect of a history of gall-bladder disease on cancer risk. Since only one discordant pair remains among those for which case and control are both positive for hypertension, the only possible estimates of relative risk in this category are RR = 0 and RR = ∞. In Chapter 7 we will see how the modelling approach, which assumes a certain structure for the joint effects of the two risk factors in each matched set, allows us to use more of the data to obtain adjusted estimates and tests for interaction between the two factors.

5.3 1:M matching: dichotomous exposures

One-to-one pair matching provides the most cost-effective design when cases and controls are equally "scarce". However when control subjects are more readily obtained than cases, which is often the case with rare forms of cancer, it may make sense to select two, three or even more controls matched to each case. According to the results of Ury (1975) (see also Breslow and Patton, 1979), the theoretical efficiency of a 1:M case-control ratio for estimating a relative risk of about one, relative to having complete information on the control population (M = ∞), is M/(M+1). Thus one control per case is 50% efficient, while four per case is 80% efficient. It is clear that increasing the number of controls beyond about 5–10 brings rapidly diminishing returns, unless one is attempting to estimate accurately an extreme relative risk.

Just as for one-to-one pair matching, we can consider each case and the corresponding controls as constituting an individual stratum. With M matched controls per case, there are 2(M+1) possible outcomes depending upon whether or not the case is exposed and upon the number of exposed controls. Each outcome corresponds to a 2×2 table.

		Exposure						
		+	−	+	−	+	−	Total
Case	+	1	0	1	0	1	0	1
Control	−	M	0	1	M−1	0	M	M
Total		M+1	0	2	M−1	1	M	M+1

(5.11)

		Exposure						
		+	−	+	−	+	−	
Case	+	0	1	0	1	0	1	1
Control	−	M	0	M−1	1	0	M	M
Total		M	1	M−1	2	0	M+1	M+1

The first and last tables have no alternative configuration, given the marginals, and hence contain no information with regard to ψ. The 2M remaining tables may be paired into sets of two, each having the same marginal total of exposed. For example, assuming $M \geqq 3$, the table with both the case and two controls positive is paired with the table with three controls positive and the case negative. More generally, we pair together the two tables

1	0	1
m–1	M–m+1	M
m	M–m+1	M+1

and

0	1	1
m	M–m	M
m	M–m+1	M+1

(5.12)

for m = 1,2, ..., M. If, as usual, p_1 denotes the probability that the case is exposed and p_0 the probability that a control is exposed, the probabilities of the two alternative outcomes in (5.12) may be written

$$\binom{M}{m-1} p_1 p_0^{m-1}(1-p_0)^{M-m+1} \text{ and } \binom{M}{m}(1-p_1)p_0^m(1-p_0)^{M-m},$$

respectively. Therefore the conditional probability of the outcome shown on the left, given the marginal totals, is

$$\text{pr(case exposed} \mid \text{m exposed among case + controls)} = \frac{m\psi}{m\psi + M-m+1}. \quad (5.13)$$

This illustrates once again the fact that consideration of the conditional distribution given the marginals eliminates the nuisance parameters and leaves the probabilities expressed solely in terms of the odds ratio ψ.

The full results of such a matched study may be summarized in the table:

		Number of controls positive				
		0	1	2	...	M
Case	Positive	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$	...	$n_{1,M}$
	Negative	$n_{0,0}$	$n_{0,1}$	$n_{0,2}$	...	$n_{0,M}$
		T_1	T_2	T_3		T_M

(5.14)

where the entry $n_{1,2}$, for example, is the number of matched sets in which the case and exactly two of the controls are exposed. The diagonal lines in (5.14) indicate the pairing of frequencies according to (5.12), i.e., $n_{1,m-1}$ with $n_{0,m}$, while the totals $T_m = n_{1,m-1} + n_{0,m}$ are the number of matched sets with exactly m subjects exposed. The conditional probability of the entire set of data may be written as a product of binomial distributions with probabilities (5.13) and is proportional to

$$\prod_{m=1}^{M} \left(\frac{m\psi}{m\psi + M-m+1}\right)^{n_{1,m-1}} \left(\frac{M-m+1}{m\psi + M-m+1}\right)^{n_{0,m}} \tag{5.15}$$

Conditional means and variances of the basic frequencies are

$$E(n_{1,m-1} \mid T_m;\psi) = \frac{T_m m\psi}{m\psi+M-m+1},$$

and (5.16)

$$Var(n_{1,m-1} \mid T_m;\psi) = \frac{T_m m\psi(M-m+1)}{(m\psi+M-m+1)^2} \quad ,$$

respectively.

Estimation

The conditional MLE i.e., the value $\hat{\psi}$ which maximizes (5.15), is obtained as the solution of the equation

$$\sum_{m=1}^{M} n_{1,m-1} = \sum_{m=1}^{M} \frac{T_m m\psi}{m\psi+M-m+1} \tag{5.17}$$

equating the total observed and expected numbers of exposed cases (see 4.25)[1]. While its solution in general requires iterative numerical calculations, a closed form expression for the case M = 2 is available (Miettinen, 1970). A more simply computed estimate is given by the robust formula (4.26) of Mantel and Haenszel, which in this case reduces to

$$\hat{\psi}_{mh} = \frac{\sum_{m=1}^{M} (M-m+1)n_{1,m-1}}{\sum_{m=1}^{M} mn_{0,m}} \quad . \tag{5.18}$$

Test of null hypothesis

As usual this is obtained by comparing the total number of exposed cases with its expectation under the null hypothesis. When $\psi = 1$ the means and variances (5.16) reduce to $T_m m/(M+1)$ and $T_m m(M-m+1)/(M+1)^2$, respectively. Hence the continuity corrected test statistic may be written

$$\chi^2 = \frac{\left\{\left|\sum_{m=1}^{M} \left(n_{1,m-1} - \frac{T_m m}{M+1}\right)\right| - {}^1/_2\right\}^2}{\frac{1}{(M+1)^2}\sum_{m=1}^{M} T_m m(M-m+1)}$$

[1] Note that the number $n_{1,M}$ of sets with the case and all the controls exposed contributes equally to both observed and expected values and is hence ignored.

$$= \frac{\left\{\left|\sum_{m=1}^{M}(M-m+1)n_{1,m-1} - \sum_{m=1}^{M} m n_{0,m}\right| - {}^1/_2(M+1)\right\}^2}{\sum_{m=1}^{M} T_m m(M-m+1)} . \qquad (5.19)$$

This is a special case of the summary chi-square formula (4.23), which has been derived both by Miettinen (1970) and Pike and Morrow (1970).

Confidence limits

Approximate confidence limits for ψ analogous to those of (4.27) are obtained from the chi-square statistic for testing hypotheses of the form $H:\psi=\psi_0$. This is similar to (5.19) but with means and variances valid for arbitrary ψ. The equations for upper and lower $100(1-\alpha)\%$ limits may thus be written

$$\frac{\sum_{m=1}^{M}\{n_{1,m-1} - E(n_{1,m-1}\,|\,T_m;\psi_L)\} - {}^1/_2}{\sqrt{\sum_{m=1}^{M} \mathrm{Var}(n_{1,m-1}\,|\,T_m;\psi_L)}} = Z_{\alpha/2}$$

and (5.20)

$$\frac{\sum_{m=1}^{M}\{n_{1,m-1} - E(n_{1,m-1}\,|\,T_m;\psi_U)\} + {}^1/_2}{\sqrt{\sum_{m=1}^{M} \mathrm{Var}(n_{1,m-1}\,|\,T_m;\psi_U)}} = -Z_{\alpha/2}$$

where E and Var are as defined in (5.16). Numerical methods are required to solve these equations.

Somewhat easier to calculate are the limits for $\log\psi$ proposed by Miettinen (1970) and based on the large sample properties of the conditional probability (5.15). According to the general theory outlined in the following chapter (§ 6.4), the approximate variance of $\log\hat{\psi}$ is

$$\mathrm{Var}(\log\hat{\psi}) = \left[\sum_{m=1}^{M} \frac{T_m m\psi(M-m+1)}{(m\psi + M - m + 1)^2}\right]^{-1} . \qquad (5.21)$$

Substituting either the MLE or M-H estimate of ψ in (5.21) to yield an estimated variance, approximate confidence limits are thus

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

i.e., (5.22)

$$\psi_L, \psi_U = \hat{\psi}\ \exp\{\pm Z_{\alpha/2}\sqrt{\mathrm{Var}\ (\log\hat{\psi})}\} .$$

Alternatively, the test-based procedure (4.20) may be used to approximate the variance, and thus the confidence limits, using only the point estimate and chi-square test

statistic. This is subject to the usual problem of underestimating the variance when ψ departs markedly from unity.

Homogeneity of the relative risk

Suppose that the matched sets have been divided into H subgroups, and that separate estimates of the odds ratio are obtained for each one. In order not to lose too much data from analysis, due to non-homogeneity of cases and controls, such subgroups are most usefully formed on the basis of variables already used for matching. The approach we shall continue to use for evaluating the statistical significance of the heterogeneity of the different estimates is to compare the observed number of exposed cases within each subgroup to that expected under the hypothesis that the same relative risk applies to all of them. Thus the statistic is a special case of that suggested in (4.32), with each matched set forming a stratum, except that the exact conditional means and variances (5.16) are used in place of the asymptotic ones.

More formally let us denote by $n_{1,m,h}$ the number of matched sets with the case and m out of M controls exposed in the h^{th} group, by $n_{0,m,h}$ the number of matched sets where the case is unexposed, and set $T_{m,h} = n_{1,m-1,h} + n_{0,m,h}$. Then the statistic for heterogeneity may be written

$$\sum_{h=1}^{H}\left[\frac{\left\{\sum_{m=1}^{M} n_{1,m-1,h} - E(n_{1,m-1,h}\,|\,T_{m,h};\hat{\psi})\right\}^2}{\sum_{m=1}^{M} \mathrm{Var}(n_{1,m-1,h}\,|\,T_{m,h};\hat{\psi})}\right] \tag{5.23}$$

where $\hat{\psi}$ is an overall estimate of the odds ratio (MLE or M-H) based on the combined data from all H subgroups. This statistic should be referred to tables of chi-square on H-1 degrees of freedom.

If the subgroups correspond to levels $x_1, \ldots, x_H$ of some quantitative variable, a single degree of freedom chi-square test for a trend in the odds ratios is obtained as

$$\frac{\left[\sum_{h=1}^{H} x_h\left\{\sum_{m=1}^{M} n_{1,m-1,h} - E(n_{1,m-1,h}\,|\,T_{m,h};\hat{\psi})\right\}\right]^2}{\sum_{h=1}^{H} x_h^2 \mathrm{Var}_h - \left[\left(\sum_{h=1}^{H} x_h \mathrm{Var}_h\right)^2 \Big/ \sum_{h=1}^{H} \mathrm{Var}_h\right]}, \tag{5.24}$$

where the $\mathrm{Var}_h = \sum_{m=1}^{M} \mathrm{Var}(n_{1,m-1,h}\,|\,T_{m,h};\hat{\psi})$ are the variances for each subgroup shown in the denominator of (5.23). Note the similarity in form between this statistic and its analog (4.31) for stratified data. When the x's are equally spaced Δ units apart, then a continuity correction of $\Delta/2$ should be applied to the numerator before squaring.

Example continued: To illustrate these methods we repeat the analyses carried out at the end of the last section, but this time we use all four controls for each case rather than just a single one. Considering as the exposure variable whether or not the subject ever used oestrogen, the basic data (5.14) are

Case		Number of controls exposed 0	1	2	3	4	Total
	Exposed	3	17	16	15	5	56
	Unexposed	0	4	1	1	1	7
		7	18	17	16		

The total number of exposed cases in the paired sets is $3+17+16+15 = 51$. According to (5.17), we find the MLE by equating this figure to its expected value,

$$51 = \frac{7\psi}{\psi+4} + \frac{36\psi}{2\psi+3} + \frac{51\psi}{3\psi+2} + \frac{64\psi}{4\psi+1}.$$

The solution $\hat{\psi} = 7.95$, obtained by numerical means, is almost identical to that calculated from the unmatched data (Table 5.1). It may be compared with the M-H estimate, determined from (5.18) as

$$\hat{\psi}_{mh} = \frac{4\times 3 + 3\times 17 + 2\times 16 + 1\times 15}{1\times 4 + 2\times 1 + 3\times 1 + 4\times 1} = 8.46.$$

The chi-square statistic (5.19) for testing H_0 is

$$\frac{\left\{|(4\times 3 + 3\times 17 + 2\times 16 + 1\times 15) - (1\times 4 + 2\times 1 + 3\times 1 + 4\times 1)| - \frac{5}{2}\right\}^2}{7\times 1\times 4 + 18\times 2\times 3 + 17\times 3\times 2 + 16\times 4\times 1}$$

$$= \frac{\left(|110-13| - \frac{1}{2}\right)^2}{302} = 29.57,$$

which is of course highly significant ($p < 0.000001$).

To obtain approximate 95% confidence limits for ψ we solve the equations (5.20)

$$\frac{51 - \left[\frac{7\psi_L}{\psi_L+4} + \frac{36\psi_L}{2\psi_L+3} + \frac{51\psi_L}{3\psi_L+2} + \frac{64\psi_L}{4\psi_L+1}\right] - {}^1/_2}{\sqrt{\frac{28\psi_L}{(\psi_L+4)^2} + \frac{108\psi_L}{(2\psi_L+3)^2} + \frac{102\psi_L}{(3\psi_L+2)^2} + \frac{64\psi_L}{(4\psi_L+1)^2}}} = 1.96$$

and

$$\frac{51 - \left[\frac{7\psi_U}{\psi_U+4} + \frac{36\psi_U}{2\psi_U+3} + \frac{51\psi_U}{3\psi_U+2} + \frac{64\psi_U}{4\psi_U+1}\right] + {}^1/_2}{\sqrt{\frac{28\psi_U}{(\psi_U+4)^2} + \frac{108\psi_U}{(2\psi_U+3)^2} + \frac{102\psi_U}{(3\psi_U+2)^2} + \frac{64\psi_U}{(4\psi_U+1)^2}}} = -1.96,$$

this requiring numerical methods, and obtain $\psi_L = 3.3$ and $\psi_U = 19.9$. It is considerably easier to calculate the variance of $\log\hat{\psi}$ using (5.20),

$$\mathrm{Var}(\log\hat{\psi}) = \left[\frac{28\hat{\psi}}{(\hat{\psi}+4)^2} + \frac{108\hat{\psi}}{(2\hat{\psi}+3)^2} + \frac{102\hat{\psi}}{(3\hat{\psi}+2)^2} + \frac{64\hat{\psi}}{(4\hat{\psi}+1)^2}\right]^{-1} = 0.177,$$

where we have inserted the MLE $\hat{\psi} = 7.95$. Consequently approximate 95% limits on $\log\psi$ are $\log(7.95) \pm 1.96\sqrt{0.177}$, or 1.249–2.899, corresponding to limits on ψ of 3.5–18.1. Finally, the test-based procedure centred about the MLE gives

$$\psi_L, \psi_U = 7.95^{(1 \pm 1.96/\sqrt{31.16})}$$

Table 5.2 Test of heterogeneity of relative risk estimates from several groups of 1:M matched sets: Los Angeles endometrial cancer study

Age group (years)	(1) Data	(2) Relative risk estimates	(3)	(4)	(5) Expectations ($\hat{\psi} = 7.95$)	(6)	(7) Variances ($\hat{\psi} = 7.95$)	(8)
	+ n_{10} n_{11} n_{12} n_{13} n_{14} − n_{00} n_{01} n_{02} n_{03} n_{04} T_1 T_2 T_3 T_4	$\hat{\psi}$	$\hat{\psi}_{mh}$	m	$E(n_{1,m-1} \mid T_m; \hat{\psi})$	Total	$Var(n_{1,m-1} \mid T_m; \hat{\psi})$	Total
55–64	0 3 4 2 2 0 1 0 0 1 1 3 4 3	4.18	4.00	1 2 3 4	0.666 2.524 3.691 2.909	9.790	0.223 0.400 0.285 0.089	0.997
65–74	1 10 10 10 2 0 1 1 1 0 2 11 11 10	9.76	10.67	1 2 3 4	1.331 9.255 10.149 9.695	30.430	0.445 1.468 0.785 0.295	2.993
75+	2 4 2 3 1 0 2 0 0 0 2 4 2 3	9.12	13.50	1 2 3 4	2.662 3.365 1.845 2.909	10.781	0.891 0.534 0.143 0.089	1.657

or limits of 3.8–16.4, which are somewhat narrower than the others as is typical of this approximate method, where the uncorrected χ^2 of 31.16 has been used, rather than the corrected value of 29.57.

Table 5.2 illustrates the procedures for evaluating the statistical significance of differences in the relative risk obtained from three different age strata. The data shown in column (1) sum to the pooled data from all three strata just analysed. We begin by calculating the means and variances of the frequencies $n_{1,m-1}$ in each stratum, under the hypothesis that the relative risk is constant across strata. Inserting the MLE $\hat{\psi} = 7.95$ in (5.16), for example, we have

$$E(n_{11} \mid T_2;\hat{\psi}) = \frac{3 \times 2 \times 7.95}{2 \times 7.95 + 3} = 2.524$$

$$\mathrm{Var}(n_{11} \mid T_2;\hat{\psi}) = \frac{3 \times 2 \times 3 \times 7.95}{(2 \times 7.95 + 3)^2} = 0.400,$$

and so on for the remaining entries in columns (5) and (7). The subtotals shown in columns (6) and (8) are the means and variances of the number of exposed cases in each stratum, excluding of course the contributions n_{1M} from matched sets in which the case and *all* controls are exposed. These quantities are inserted in (5.23) to obtain the test statistic, with two degrees of freedom

$$\chi^2 = \frac{(9-9.790)^2}{0.997} + \frac{(31-30.430)^2}{2.993} + \frac{(11-10.781)^2}{1.657} = 0.76.$$

Hence, there is no evidence ($p = 0.68$) of heterogeneity, the variations between the stratum-specific relative risk estimates shown in columns (2) and (3) being attributable to the small numbers in each table.

For the sake of completeness we compute also the single degree of freedom chi-square (5.24) for a trend in relative risk with age, although we know already that its value cannot exceed the 0.76 just obtained for the overall comparison. Assigning "doses" of $x_1 = 0$, $x_2 = 1$ and $x_3 = 2$ to the three age strata, we have

$$\chi^2 = \frac{\{(31-30.430) + 2(11-10.781) - {}^1/_2\}^2}{2.993 + 4 \times 1.657 - (2.993 + 2 \times 1.657)^2/(0.997 + 2.993 + 1.657)} = 0{,}09,$$

where a continuity correction of $^1/_2$ is applied to the numerator in view of the fact that the x's are spaced one unit apart.

5.4 Dichotomous exposure: variable number of controls

Although the study design stipulates that a fixed number of controls be matched to each case, in practice it may not always be possible to locate the full complement of controls. Even for sets in which all controls are available, some may lack information regarding certain of the risk factors. If the original design calls for 1:4 matching, for example, one may end up with most of the matched sets having data on 1 case and 4 controls, while a lesser number have 3, 2 or 1 controls. Of course, sets in which data are available only for the case, or only for the controls, provide no information about the relative risk in a matched analysis and hence need not be considered.

One approach to the analysis of matched sets containing a variable number of controls is simply to discard all those which do not contain the full number specified by design. Clearly this is a waste of important information and would be considered only if the number of sets to be discarded represented a small fraction of the total. A slight information loss might then be tolerated in order not to increase the computational burden.

Fortunately, the extra computation required is not that great. All of the tests and estimates considered in the previous section may be broken down into component

parts consisting of sums or linear combinations of the observed frequencies (5.14), their means and their variances. The corresponding statistic may be generalized for use with a variable number of controls simply by computing each component part separately for the matched sets having a specified case-control ratio, and then reassembling the parts.

Arranging the data as in Table 5.3, let $n_{i,m,M}$ denote the number of matched sets containing M controls of which m are exposed and the case is (i = 1) or is not (i = 0) exposed. Let $T_{m,M} = n_{1,m-1,M}+n_{0,m,M}$ denote the number of such sets having a total of m exposed. The M-H estimate of relative risk may then be written

$$\hat{\psi}_{mh} = \frac{\sum_{M}\sum_{m=1}^{M}(M-m+1)n_{1,m-1,M}/(M+1)}{\sum_{M}\sum_{m=1}^{M} mn_{0,m,M}/(M+1)} \tag{5.25}$$

where $\sum_{M}$ denotes summation over the data in the sub-tables formed for each case-control ratio. The MLE is found as before by equating the observed and expected numbers of exposed cases, as in (5.17), except that there will be a separate contribution to the left and right hand sides of the equation for each value of M:

$$\sum_{M}\sum_{m=1}^{M} n_{1,m-1,M} = \sum_{M}\sum_{m=1}^{M} \frac{T_{m,M}m\psi}{(m\psi+M-m+1)} \quad . \tag{5.26}$$

Similarly, the statistic (5.19) for testing the null hypothesis may be written in terms of separate contributions to the observed and expected values, as well as the variance, from each sub-table:

$$\chi^2 = \frac{\left[\left|\sum_{M}\sum_{m=1}^{M}\left(n_{1,m-1,M} - \frac{m}{M+1}T_{m,M}\right)\right| - {}^1/_2\right]^2}{\sum_{M}\sum_{m=1}^{M} T_{m,M}m(M-m+1)/(M+1)^2} . \tag{5.27}$$

Corresponding adjustments are made to the equations (5.20) and (5.21) used to find confidence intervals, as well as to the statistics (5.23) and (5.24) used to test the heterogeneity of the odds ratio in different strata.

Kodlin and McCarthy (1978) note that the M-H estimate (5.25) and summary chi-square (5.27) may each be represented in terms of weighted sums of the basic data appearing in Table 5.3. Appropriate coefficients for weighting each entry are shown in Table 5.4, of which the five parts correspond, respectively, to the numerator and denominator of the M-H estimate, the observed and expected numbers of exposed cases (excluding sets where the case and all controls are exposed), and the variance of the number of exposed cases. For example, using Part A of the table the numerator of the M-H statistic would be calculated as

$$\frac{1}{2}n_{1,0,1}$$
$$+\frac{2}{3}n_{1,0,2}+\frac{1}{3}n_{1,1,2}$$
$$+\frac{3}{4}n_{1,0,3}+\frac{2}{4}n_{1,1,3}+\frac{1}{4}n_{1,2,3}$$
$$\vdots$$
$$+\frac{M}{M+1}n_{1,0,M}+\frac{M-1}{M+1}n_{1,1,M}+\frac{M-2}{M+1}n_{1,2,M}+\ldots+\frac{1}{M+1}n_{1,M-1,M}.$$

Table 5.3 Data layout for a matched study involving variable number of controls

Case : control ratio	Exposure of case	Number of controls exposed 0	1	2	...	M
1:1	+	$n_{1,0,1}$	$n_{1,1,1}$			
	−	$n_{0,0,1}$	$n_{0,1,1}$			
	+	$n_{1,0,2}$	$n_{1,1,2}$	$n_{1,2,2}$		
1:2	−	$n_{0,0,2}$	$n_{0,1,2}$	$n_{0,2,2}$		
.	.	.	.			
.	.	.	.			
.	.	.	.			
1:M	+	$n_{1,0,M}$	$n_{1,1,M}$	$n_{1,2,M}$	...	$n_{1,M,M}$
	−	$n_{0,0,M}$	$n_{0,1,M}$	$n_{0,2,M}$		$n_{0,M,M}$

Example continued: To illustrate the procedure to be followed with variable numbers of controls per case we selected another risk variable, dose of *conjugated* oestrogen, for which several subjects had missing values (Table 5.1). Four matched sets in which the case had a missing value were excluded from this analysis. The 59 remaining sets could be divided into two categories, 55 having 4 controls and 4 having 3 controls. Thus, defining "exposed" to be anything above a zero dose of conjugated oestrogen, the results were summarized:

Case: control ratio	Exposure for case	Number of controls exposed 0	1	2	3	4	Total
1:3	+	1	3	0	0		4
	−	0	0	0	0		0
		1	3	0			
1:4	+	4	17	11	9	2	43
	−	1	6	3	1	1	12
		10	20	12	10		

Table 5.4 Coefficients used for weighted sums in calculation of the M–H estimate and summary chi-square from matched sets with variable numbers of controls[a]

Case : control ratio	Case exposure	Number of controls exposed					
		0	1	2	3	...	M
		A. Numerator of M–H estimate					
1:1	+	$^1/_2$	0				
1:2	+	$^2/_3$	$^1/_3$	0			
1:3	+	$^3/_4$	$^2/_4$	$^1/_4$	0		
.				.			
.				.			
.				.			
1:M	+	$\frac{M}{M+1}$	$\frac{M-1}{M+1}$	$\frac{M-2}{M+1}$	$\frac{M-3}{M+1}$	...	0
		B. Denominator of M–H estimate					
1:1	–	0	$^1/_2$				
1:2	–	0	$^1/_3$	$^2/_3$			
1:3	–	0	$^1/_4$	$^2/_4$	$^3/_4$		
.	.			.			
.	.			.			
.	.			.			
1:M	–	0	$\frac{1}{M+1}$	$\frac{2}{M+1}$	$\frac{3}{M+1}$	...	$\frac{M}{M+1}$

C. Observed number of exposed cases

Case : control ratio	Case exposure	Number of controls exposed 0	1	2	3	...	M
1:1	+	0	1				
1:2	+	0	1	1			
1:3	+	0	1	1	1		
.	.			.			
.	.			.			
.	.			.			
1:M	+	0	1	1	1	...	1

D. Expected number of exposed cases (H_0)

1:1	+	$^1/_2$	0				
	−	0	$^1/_2$				
1:2	+	$^1/_3$	$^2/_3$	0			
	−	0	$^1/_3$	$^2/_3$			
1:3	+	$^1/_4$	$^2/_4$	$^3/_4$	0		
	−	0	$^1/_4$	$^2/_4$	$^3/_4$		
.	.			.			
.	.			.			
.	.			.			
1:M	+	$\frac{1}{M+1}$	$\frac{2}{M+1}$	$\frac{3}{M+1}$	$\frac{4}{M+1}$	...	0
	−	0	$\frac{1}{M+1}$	$\frac{2}{M+1}$	$\frac{4}{M+1}$		$\frac{M}{M+1}$

E. Variance of numbers of exposed cases (H_0)

Case : control ratio	Case exposure	Number of controls exposed 0	1	2	3	...	M
1 : 1	+	$^1/_4$	0				
	−	0	$^1/_4$				
1 : 2	+	$^2/_9$	$^2/_9$	0			
	−	0	$^2/_9$	$^2/_9$			
1 : 3	+	$^3/_{16}$	$^4/_{16}$	$^3/_{16}$	0		
	−	0	$^3/_{16}$	$^4/_{16}$	$^3/_{16}$		
.	.			.			
.	.			.			
.	.			.			
1 : M	+	$\frac{M}{(M+1)^2}$	$\frac{2(M-1)}{(M+1)^2}$	$\frac{3(M-2)}{(M+1)^2}$	$\frac{4(M-3)}{(M+1)^2}$	...	0
	−	0	$\frac{M}{(M+1)^2}$	$\frac{2(M-1)}{(M+1)^2}$	$\frac{3(M-2)}{(M+1)^2}$		$\frac{M}{(M+1)^2}$

[a] When parts of the data are not shown, the corresponding coefficients are zero.

Accordingly, the M-H estimate, calculated from (5.25), is

$$\hat{\psi}_{mh} = \frac{\frac{3\times1}{4} + \frac{2\times3}{4} + \frac{4\times4}{5} + \frac{3\times17}{5} + \frac{2\times11}{5} + \frac{1\times9}{5}}{\frac{1\times6}{5} + \frac{2\times3}{5} + \frac{3\times1}{5} + \frac{4\times1}{5}} = \frac{21.85}{3.80} = 5.75,$$

while the equation (5.26) to be solved for the MLE is

$$45 = 1 + 3 + 4 + 17 + 11 + 9$$

$$= \frac{\psi}{\psi+3} + \frac{3\times2\psi}{2\psi+2} + \frac{10\times\psi}{\psi+4} + \frac{20\times2\psi}{2\psi+3} + \frac{12\times3\psi}{3\psi+2} + \frac{10\times4\psi}{4\psi+1},$$

yielding $\hat{\psi} = 5.53$. To test the null hypothesis we first find the mean value

$$\sum_{M}\sum_{m=1}^{M}\frac{mT_{m,M}}{M+1} = \frac{1\times1}{4} + \frac{2\times3}{4} + \frac{1\times10}{5} + \frac{2\times20}{5} + \frac{3\times12}{5} + \frac{4\times10}{5} = 26.95$$

and the variance

$$\sum_{M}\sum_{m=1}^{M}\frac{m(M-m+1)T_{m,M}}{(M+1)^2}=\frac{1\times3\times1}{16}+\frac{2\times2\times3}{16}+\frac{1\times4\times10}{25}+\frac{2\times3\times20}{25}+\frac{3\times2\times12}{25}+\frac{4\times1\times10}{25}$$

$$=11.82,$$

from which the test statistic (5.27) is

$$\chi^2=\frac{(45-26.95-{}^1/_2)^2}{11.82}=26.06.$$

Ninety-five percent confidence limits for $\log\psi$ based on (5.21) are found by calculating the variance with separate contributions for M = 3 and M = 4:

$$\text{Var}=\left[1\frac{\psi\times3}{(\psi+3)^2}+3\frac{2\psi\times2}{(2\psi+2)^2}+10\frac{\psi\times4}{(\psi+4)^2}+20\frac{2\psi\times3}{(2\psi+3)^2}+12\frac{3\psi\times2}{(3\psi+2)^2}+10\frac{4\psi\times1}{(4\psi+1)^2}\right]^{-1}=0.125,$$

where we have inserted the MLE for ψ. Consequently the confidence limits are

$$\psi_L,\psi_U=5.53\times\exp(\pm1.96\sqrt{0.125})$$
$$=(2.76,\ 11.1).$$

5.5 Multiple exposure levels: single control

Restriction of a risk variable to two levels may waste important information about the effects of the full range of exposures actually experienced (§ 4.5). More detailed results are obtained if the case and control in each matched pair are classified instead into one of several exposure categories. The data are usefully summarized as in Table 5.5, where the entry n_{kh} denotes the number of pairs in which the case is exposed at level k and the control at level h of K possible levels. The marginal totals $n_{k.}$ and $n_{.k}$ represent, respectively, the total number of cases and total number of controls which are exposed at level k. This situation has been studied in some detail by Pike, Casagrande and Smith (1975).

Table 5.5 Representation of data from a matched pair study with K exposure categories

Exposure level for case	Exposure level for control				
	1	2	...	K	Total
1	n_{11}	n_{12}	...	n_{1K}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2K}	$n_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
K	n_{K1}	n_{K2}	...	n_{KK}	$n_{K.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.K}$	$n_{..}$

Following the general principles of conditional inference outlined in § 4.2 and § 4.3, we approach the analysis of such data by considering the probability of the outcome in each matched pair conditional on the combined set of exposures for case and control. Pairs in which both members are exposed to the same level are uninformative about the relative risk since for them the conditional probability of the observed outcome is unity. Hence the statistical analysis does not utilize the diagonal entries n_{kk} in Table 5.5. The off-diagonal entries in the table may be grouped into sets of two representing all pairs having a particular combination of different exposures. Thus, for $k \neq h$, $N_{kh} = n_{kh} + n_{hk}$ represents the number of matched pairs in which the exposures are at levels k and h, without specifying which is associated with the case and which with the control.. If ψ_{kh} denotes the relative risk of disease for level k *versus* that for level h, then the conditional distribution of n_{kh} given N_{kh} is binominal (cf. 5.3):

$$\text{pr}(n_{kh} \mid N_{kh}) = \binom{N_{kh}}{n_{kh}} \left(\frac{\psi_{kh}}{1+\psi_{kh}}\right)^{n_{kh}} \left(\frac{1}{1+\psi_{kh}}\right)^{n_{hk}} . \tag{5.28}$$

The (conditional) distribution of the entire set of data consists of the product of K(K–1)/2 such binomials, one for each of the entries n_{kh} above the diagonal ($k<h$) in Table 5.5.

Estimation of relative risk

As noted in § 4.5 for the combination of multiple exposure level data across several strata, the summary estimates of relative risk for different pairs of exposure levels may not display the consistency expected of them. The same phenomenon occurs with matched pairs. Here the odds ratio relating levels k and h of exposure may be calculated from the pairs showing exposure to these two levels only (cf. 5.6) as the ratio

$$\hat{\psi}_{kh} = \frac{n_{kh}}{n_{hk}} .$$

According to their interpretation as ratios of incidence rates for level k *versus* level h, assumed to be constant across the matching factors, the estimated odds ratios ought to satisfy, within the bounds of sampling error, the consistency relationship

$$\psi_{kh} = \frac{\psi_k}{\psi_h} , \tag{5.29}$$

where $\psi_k = \psi_{k1}$ and $\psi_h = \psi_{h1}$ denote the odds ratios for levels k and h relative to level 1 (baseline). To the extent that the individual estimates $\hat{\psi}_{kh}$ do not satisfy this condition, at least within the limits of random variation, the assumption of constant relative risks across the factors used for matching is called into question.

In order to ensure that the estimated relative risks do display such consistency it is necessary to *build the relationship into a model* for the observed data. The model will contain K–1 parameters $\psi_2, \ldots, \psi_K$ whose ratios are assumed to represent the relative risks for each pair of levels as in (5.29). It is an example of the general conditional model for matched data which will be discussed at greater length in Chapter 7. MLEs for the parameters in the model are found from the usual set of formulae equating

observed and expected values of the numbers of cases exposed to each level. There are K–1 equations in K–1 unknowns, namely[1]

$$\sum_{h:h\neq k} n_{kh} = \sum_{h:h\neq k} N_{kh} \frac{\psi_k}{\psi_k+\psi_h} \tag{5.30}$$

for k = 2, ..., K. Solution requires numerical methods. Variances for the estimates are also available, but discussion of their derivation and computation is perhaps best left until presentation of the general model (§ 7.3). Approximate confidence limits for the parameters ψ_k may be based on these variances.

Test of the null hypothesis

A test of the hypothesis $H_0: \psi_2 = \psi_3 = \ldots = \psi_K = 1$ that there is no effect of exposure on risk is obtained by comparing the observed numbers of cases exposed at each level to that expected, standardizing by the corresponding variance-covariance matrix. Since all the probabilities $\psi_k/(\psi_k+\psi_h)$ in (5.30) are equal to $^1/_2$ under H_0, the means, variances and covariances of the marginal totals shown in Table 5.5 are readily calculated to be

$$\begin{aligned} E(n_{k.}) &= {}^1/_2\,(n_{k.}+n_{.k}) \\ Var(n_{k.}) &= {}^1/_4\,(n_{k.}+n_{.k}) - {}^1/_2\,n_{kk} \\ \text{and} \quad Cov(n_{k.},n_{h.}) &= -{}^1/_4\,N_{kh}, \text{ for } h \neq k., \end{aligned} \tag{5.31}$$

respectively. Only the first K–1 of these are used to form the test statistic, defined by

$$(\mathbf{O}-\mathbf{E})^T \mathbf{V}^{-1} (\mathbf{O}-\mathbf{E}) \tag{5.32}$$

where **O** and **E** denote the K–1 dimensional vectors of observed and expected values of the $n_{k.}$, while **V** is the corresponding (K–1) × (K–1) dimensional covariance matrix. This has a nominal χ^2_{K-1} distribution under the null hypothesis. First proposed by Stuart (1955), it is a special case of the general summary chi-square (4.41) used for testing homogeneity with stratified data (Mantel & Byar, 1978).

If dose levels $x_1, \ldots, x_K$ are assigned to the K exposure levels, a test for a linear trend in the (log) relative risks ψ_k with increasing dose may be based on the statistic[2]

$$\frac{\left\{\sum\sum_{k<h} (n_{kh}-n_{hk})(x_k-x_h)\right\}^2}{\sum\sum_{k<h} N_{kh}(x_k-x_h)^2}. \tag{5.33}$$

To make a continuity correction the absolute value of the numerator term inside the brackets is reduced by half of the difference between adjacent doses, provided these

[1] Here $\sum_{h:h\neq k}$ means summation over the indices h which are not equal to a fixed k, i.e., $\sum_{h:h\neq 3} n_{3h} = n_{31} + n_{32} + n_{34} + \ldots$

[2] Here and below $\sum\sum_{k<h}$ denotes summation over all K(K–1)/2 *pairs* of indices (k, h) with k<h.

are equally spaced. This statistic, a special case of (4.43), should be referred to tables of chi-square with one degree of freedom.

Testing for consistency of the odds ratio

In order to test for consistency in the estimated odds ratios, which as explained earlier (§ 4.5) is a consequence of our usual assumptions about the constancy of the relative risk, we compare the frequencies observed in Table 5.5 with those expected under the hypothesis (5.29) using the usual chi-square formula. More specifically, the test statistic is defined by

$$\sum_{k<h}\sum \left\{ n_{kh} - N_{kh} \frac{\hat{\psi}_k}{\hat{\psi}_k + \hat{\psi}_h} \right\}^2 \times \frac{(\hat{\psi}_k + \hat{\psi}_h)^2}{N_{kh}\hat{\psi}_k\hat{\psi}_h}$$

$$= \sum_{k<h}\sum \frac{(n_{kh}\hat{\psi}_h - n_{hk}\hat{\psi}_k)^2}{N_{kh}\hat{\psi}_k\hat{\psi}_h}, \qquad (5.34)$$

where the $\hat{\psi}_k$ are the ML estimates obtained from (5.30).

This statistic should be referred to tables of chi-square with K(K–1)/2–(K–1) = (K–1)(K–2)/2 degrees of freedom. A significant result would lead one to reject the hypothesis of consistency and to search for matching variables which modified the relative risks. However, this test is not likely to be as sensitive to such interactions as the more direct methods based on the modelling approach.

Example continued: We have already remarked that for 4 of 63 cases from the Los Angeles endometrial cancer study the dose level of conjugated oestrogen was unknown. However, this variable was known for the first matched control in all sets. Using four levels of exposure, (1) none, (2) 0.1–0.299 mg, (3) 0.3–0.625 mg and (4) 0.626+ mg, the data for the 59 matched pairs are presented in Table 5.6.

To estimate the relative risk parameters ψ_2, ψ_3, ψ_4, for levels 2, 3 and 4 *versus* level 1, assuming consistency, we set up the equations (5.30):

Table 5.6 Average doses of conjugated oestrogen used by cases and matched controls: Los Angeles endometrial cancer study

Average dose for case (mg)	Average dose for control (mg)				
	0	0.1–0.299	0.3–0.625	0.626+	Total
0	6	2	3	1	12
0.1–0.299	9	4	2	1	16
0.3–0.625	9	2	3	1	15
0.626+	12	1	2	1	16
Total	36	9	10	4	59

$$9 + 2 + 1 = 11\frac{\psi_2}{1+\psi_2} + 4\frac{\psi_2}{\psi_2+\psi_3} + 2\frac{\psi_2}{\psi_2+\psi_4}$$

$$9 + 2 + 1 = 12\frac{\psi_3}{1+\psi_3} + 4\frac{\psi_3}{\psi_2+\psi_3} + 3\frac{\psi_3}{\psi_3+\psi_4}$$

$$12 + 1 + 2 = 13\frac{\psi_4}{1+\psi_4} + 2\frac{\psi_4}{\psi_2+\psi_4} + 3\frac{\psi_4}{\psi_3+\psi_4} .$$

Their solution, obtained by numerical methods, is $\hat{\psi}_2 = 4.59$, $\hat{\psi}_3 = 3.55$ and $\hat{\psi}_4 = 8.33$. These values may be inserted in (5.34) to test the assumption of consistency, yielding

$$\frac{(2\times 4.59-9\times 1)^2}{11\times 1\times 4.59} + \frac{(3\times 3.55-9\times 1)^2}{12\times 1\times 3.55} + \frac{(1\times 8.33-12\times 1)^2}{13\times 1\times 8.33}$$

$$+ \frac{(2\times 3.55-2\times 4.59)^2}{4\times 4.59\times 3.55} + \frac{(1\times 8.33-1\times 4.59)^2}{2\times 4.59\times 8.33} + \frac{(1\times 8.33-2\times 3.55)^2}{3\times 3.55\times 8.33} = 0.46,$$

which when referred to tables of chi-square with $(4-1)(4-2)/2 = 3$ degrees of freedom gives $p = 0.93$. In other words, the observed data satisfy the consistency hypothesis extremely well.

In order to carry out the global test of the null hypothesis we calculate the means

$$\begin{aligned} E(n_{1.}) &= {}^1/_2(36+12) = 24 \\ E(n_{2.}) &= {}^1/_2(\ 9+16) = 12.5 \\ E(n_{3.}) &= {}^1/_2(10+15) = 12.5, \end{aligned}$$

variances

$$\begin{aligned} \mathrm{Var}(n_{1.}) &= {}^1/_4(36+12)-{}^1/_26 = 9 \\ \mathrm{Var}(n_{2.}) &= {}^1/_4(9+16)-{}^1/_24 = 4.25 \\ \mathrm{Var}(n_{3.}) &= {}^1/_4(10+15)-{}^1/_23 = 4.75 \end{aligned}$$

and covariances

$$\begin{aligned} \mathrm{Cov}(n_{1.},n_{2.}) &= -{}^1/_4(2+9) = -2.75 \\ \mathrm{Cov}(n_{1.},n_{3.}) &= -{}^1/_4(3+9) = -3 \\ \mathrm{Cov}(n_{2.},n_{3.}) &= -{}^1/_4(2+2) = -1 \end{aligned}$$

according to (5.31). The test statistic (5.32) is then

$$(36-24,9-12.5,10-12.5)\begin{bmatrix} 9 & -2.75 & -3 \\ -2.75 & 4.25 & -1 \\ -3 & -1 & 4.75 \end{bmatrix}^{-1}\begin{pmatrix} 36-24 \\ 9-12.5 \\ 10-12.5 \end{pmatrix}$$

$$= (12,-3.5,-2.5)\begin{bmatrix} 0.234 & 0.196 & 0.189 \\ 0.196 & 0.412 & 0.210 \\ 0.189 & 0.210 & 0.374 \end{bmatrix}\begin{pmatrix} 12 \\ -3.5 \\ -2.5 \end{pmatrix}$$

$$= 16.96,$$

which is highly significant ($p = 0.001$) as shown by reference to tables of chi-square with three degrees of freedom. Assigning dose levels of $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 4$ to the four exposure levels, we next calculate the test for trend using (5.33). This is

$$\frac{[(2-9)(1-2)+(3-9)(1-3)+(1-12)(1-4)+(2-2)(2-3)+(1-1)(2-4)+(1-2)(3-4)-{}^1/_2]^2}{11(1-2)^2+12(1-3)^2+13(1-4)^2+4(2-3)^2+2(2-4)^2+3(3-4)^2} = 14.43,$$

an even more significant result ($p = 0.0001$) which indicates that most of the variation in risk among the four exposure levels is accounted for by the linear increase. The contribution of $16.96-14.43 = 2.53$

from the remaining two degrees of freedom is not statistically significant. Note that we have used the continuity correction of $^1/_2$ in the numerator of this statistic, as is appropriate since the assigned x's are spaced one unit apart.

5.6 More complex situations

One lesson learned from the preceding sections is that the types of matched data which can be analysed easily using elementary methods are extremely limited. While the calculations are reasonably tractable in the case of a single dichotomous risk variable, with both single or multiple controls, estimation of a consistent set of relative risks for polytomous exposures requires solution of a system of non-linear equations even for matched pairs. More complicated still are the situations involving multiple controls together with a single exposure variable at multiple levels, or multiple exposure variables with any combination of controls. The control of confounding, or evaluation of effect modification, by variables not used for matching may require that we discard from analysis much of the relevant data.

Certain of the limitations imposed by the elementary methods can be overcome using multivariate analysis. Just as we noted earlier for stratified samples, multivariate analysis of matched data is carried out in the context of an explicit mathematical model relating each individual's exposures to his risk for disease. Such modelling is especially valuable in dealing with quantitative variables as it permits their effect on risk to be summarized by a few parameters. Chapter 6 introduces for this purpose the linear logistic regression model, showing that its structure is well suited for determining the multiplicative effects of one or more risk factors on disease rates. Chapter 7 extends the model for use with matched or finely stratified samples. Since all the tests and estimates considered in this chapter occur as special cases of those derived from the general model, the general-purpose computer programmes (Appendix IV) which are available to fit the multivariate model can be used (and in fact were used) to solve the equations for maximum likelihood estimation which occur in those particular problems considered above.

REFERENCES

Breslow, N. & Patton, J. (1979) *Case-control analysis of cohort studies.* In Breslow, N. & Whittemore, A., eds, *Energy and Health,* Philadelphia. Society for Industrial and Applied Mathematics, pp. 226–242

Kodlin, D. & McCarthy, N. (1978) Reserpine and breast cancer. *Cancer, 41,* 761–768

Mack, T.M., Pike, M.C., Henderson, B.E., Pfeffer, R.I., Gerkins, V.R., Arthur, B.S. & Brown, S.E. (1976) Estrogens and endometrial cancer in a retirement community. *New Engl. J. Med., 294,* 1262–1267

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12,* 153–157

Mantel, N. & Byar, D. (1978) Marginal homogeneity, symmetry and independence. *Commun. Stat. Theory Meth., A7 (10),* 956–976

Miettinen, O.S. (1970) Estimation of relative risk from individually matched series. *Biometrics, 26,* 75–86

Pearson E.S. & Hartley, H.O. (1966) *Biometrika Tables for Statisticians,* Vol. I (3rd Edition), Cambridge, Cambridge University Press

Pike, M.C., Casagrande, J. & Smith, P.G. (1975) Statistical analysis of individually matched case-control studies in epidemiology: factor under study a discrete variable taking multiple values. *Br. J. prev. soc. Med., 29,* 196–201

Pike, M.C. & Morrow, R.H. (1970) Statistical analysis of patient-control studies in epidemiology. Factor under investigation an all-or-none variable. *Br. J. prev. soc. Med., 24,* 42–44

Stuart, A. (1955) A test for homogeneity of the marginal distributions in a two way classification. *Biometrika, 42,* 412–416

Ury, H.K. (1975) Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics, 31,* 643–649

LIST OF SYMBOLS – CHAPTER 5 (in order of appearance)

n_{11}	number of matched pairs with both case and control exposed
n_{10}	number of matched pairs with case exposed and control not
n_{01}	number of matched pairs with control exposed and case not
n_{00}	number of matched pairs with neither case nor control exposed
p_1	probability of exposure for case
$q_1 = 1-p_1$	probability of non-exposure for case
p_0	probability of exposure for control
$q_0 = 1-p_0$	probability of non-exposure for control
ψ	odds ratio
π	probability that in a discordant matched pair it is the case who is exposed rather than the control
E( )	expectation of a quantity ( )
Var( )	variance of a quantity ( )
$\binom{n_1+n_2}{n_1}$	binomial coefficient; number of ways of drawing samples of n_1 objects from a total of n_1+n_2
$\|x\|$	absolute value of a number x
π_L	lower confidence limit for π
π_U	upper confidence limit for π
ψ_L	lower confidence limit for ψ
ψ_U	upper confidence limit for ψ
$Z_{\alpha/2}$	upper $100\alpha/2$ percentile of the standard normal distribution
pr( )	probability of an event ( )
pr(|)	probability of one event conditional on another
M	number of controls in each matched set
$n_{1,m}$	number of matched sets with case exposed and m controls exposed
$n_{0,m}$	number of matched sets with case not exposed and m controls exposed
T_m	number of matched sets with m exposed among case + controls (additional subscripts are added to distinguish various groups)

$E(\mid)$	expectation of a quantity conditional on the values of another
$Var(\mid)$	variance of a quantity conditional on the values of another
h	subscript indicating the h^{th} of H groups of matched sets; e.g., $n_{1,m,h}$ is the number of matched sets with the case and m controls exposed in the h^{th} group
M	subscript indicating the number of controls in matched set data having a variable number of controls per case; e.g., $n_{1,m,M}$ is the number of sets in which the case and m of M controls are exposed
χ^2_ν	a statistic whose (asymptotic) distribution under the null hypothesis is that of chi-square on ν degrees of freedom (when ν is not specified it is meant to be 1)
K	number of levels of a polytomous risk factor
n_{kh}	number of matched pairs where the case is exposed at level k of a polytomous variable and the control at level h
$N_{kh} = n_{kh} + n_{hk}$	number of matched pairs in which one member is at level k and the other at level h ($k \neq h$)
$n_{k.}$	sum of n_{kh} over h; number of matched pairs where the case is exposed at level k
$n_{.k}$	sum of n_{hk} over h; number of matched pairs where the control is exposed at level k
ψ_{kh}	odds ratio expressing relative risk for exposure to level k *versus* level h of a polytomous variable
ψ_k	odds ratio expressing relative risk of disease for a person exposed to level k of a polytomous factor, using level 1 as baseline ($\psi_1 = 1$)
$\hat{}$	denotes an estimate, e.g., $\hat{\psi}_k$ is an estimate of the odds ratio ψ_k
O	K–1 dimensional vector of numbers of matched pairs in which the case is exposed to one of the first K–1 levels of a polytomous factor; $\mathbf{O} = (n_{1.}, n_{2.}, \ldots, n_{K-1.})$
E	K–1 vector of expectations $\mathbf{E} = E(\mathbf{O}) = [E(n_{1.}), E(n_{2.}), \ldots, E(n_{K-1,.})]$
V	K–1 × K–1 variance-covariance matrix of which the (k,h) element is $Cov(n_{k.}, n_{h.}) = -{}^1/_4 N_{kh}$ for $k \neq h$ or $Var(n_{k.}) = {}^1/_4(n_{k.} + n_{.k}) - {}^1/_2 n_{kk}$ for k=h (see equation 5.3)