19
Individually matched case-control studies

Analyses which preserve the matching of individual cases to their controls follow similar principles to those of Chapter 18. The strata are now the sets made up of each case and its matched controls. Studies designed to have a fixed number of controls, m say, drawn for each case, will be referred to as 1:m matched studies.

19.1 Mantel–Haenszel analysis of the 1:1 matched study

For reasons discussed in Chapter 18, the use of profile likelihood gives misleading estimates of odds ratios when there are a large number of strata with little data in each stratum. However, the Mantel–Haenszel method works perfectly well in these circumstances. The calculations are particularly easy in the 1:1 case, and illustrate ideas which are important for our later discussion of the likelihood approach.

The results of 1:1 matched studies are usually presented in 2×2 tables such as Table 19.1.* These data were drawn from the same study as reported in Chapter 17, and concern the relationship between tonsillectomy history and the incidence of Hodgkin's disease. The total study included 174 cases and 472 controls, but the controls were siblings of the cases, and the authors felt that the matching of cases and sibling controls should be preserved. They also wished to control for age and sex and therefore restricted their analysis to 85 matched case-control pairs in which the case and sibling control were of the same sex and matched for age within a specified margin. Note that, in the construction of matched sets, the original 174 cases and 472 controls have been reduced to only 85 cases and 85 controls.

Tables such as Table 19.1 can be confusing because we are used to seeing tables that count subjects, while this table counts case-control sets. The four cells of the table correspond to the four possible exposure configurations of a case-control set. These are illustrated in terms of a tree in Fig. 19.1. The first branching point is according to whether or not the control was exposed (denoted E+ and E- respectively), while the second

*From Cole, P. *et al.* (1973) *New England Journal of Medicine*, **288**, 634.

Table 19.1. Tonsillectomy history in 85 matched pairs

History of case	History of control	
	Positive	Negative
Positive	26	15
Negative	7	37

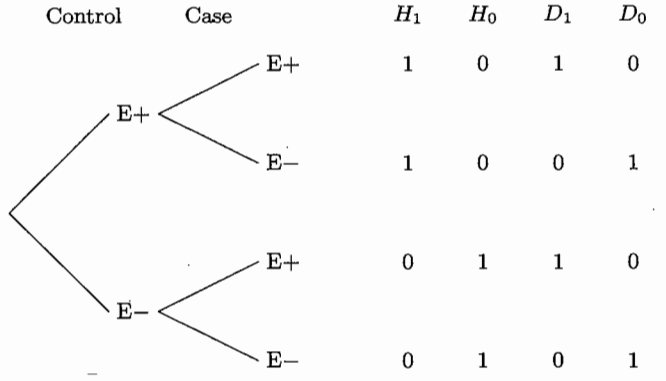

Fig. 19.1. Exposure configurations for 1:1 sets.

branching is according to exposure of the case. The frequencies in Table 19.1 refer to counts of these four configurations.

Exercise 19.1. How often did each of the exposure configurations of Fig. 19.1 occur?

In the analysis of individually matched studies the strata are case-control sets so that, in the notation of Chapter 18, t indexes sets. The number of subjects in each stratum is $N^t = 2$, and since each stratum contains one case and one control, D^t and H^t are always 1. The values of D_1^t, D_0^t, H_1^t, and H_0^t for each exposure configuration are shown in Fig. 19.1. In this figure and henceforth we will omit the superscript t for clarity, and remember that the symbols refer to values in a single case-control set.

Exercise 19.2. What are the contributions of each configuration to Q and R in the Mantel–Haenszel estimate of the odds ratio? Similarly what are the contributions to the score and score variance, U and V? Which configurations contribute to estimation and testing?

It can be seen that only two exposure configurations make any contribution to estimation and testing of the odds ratio. These are the sets in which the exposure status of case and controls differ and are called *discordant* sets. The remaining sets are called *concordant sets*. In our current example, 63

of the case-control sets are concordant and are ignored.

Exercise 19.3. For the tonsillectomy data, what are the values for Q, R, U, V? Using the methods of Chapter 18, estimate the odds ratio, its 90% confidence interval, and a p-value for $\theta = 1$.

The odds ratio estimate is very close to that obtained in the analysis of Chapter 17, but so much data has been lost in this analysis that the result is no longer statistically significant. It is easy to criticize an analysis which discards so much data, but when it is necessary to preserve the matching of controls to cases it is not easy to see how one can adjust for the effects of additional variables by stratification, since the case and its control may fall within different strata. At the time this study was reported there would have been no alternative but to discard such sets. Nowadays, this problem is easily overcome by use of the regression methods to be described in Part II.

Before leaving this example, it is interesting to note that the above analysis is not the one originally reported. In their first report, the researchers subscribed to the misconception discussed in Chapter 18 — that the matching for age, sex, and family was sufficient to control for these variables and that subsequently the matching could be ignored in the analysis.

Exercise 19.4. Show that the odds ratio estimate obtained by ignoring the matching is less than that obtained by the correct analysis.

19.2 The hypergeometric likelihood for 1:1 matched studies

The hypergeometric likelihood is obtained by arguing conditionally upon *both* margins of the 2×2 table, and depends only upon the odds ratio parameter. It is usually difficult to compute, but its use is only necessary when the data within strata are few. This is the case for individually matched studies and the hypergeometric likelihood *must* be used. Luckily in this case the computations are quite easy — particularly in the 1:1 case.

Fig. 19.2 derives the probability of each exposure configuration by multiplying along branches of the tree in the usual way and also lists the total number of *subjects* in the set who were exposed, N_1. The odds that the control in the set was exposed is denoted by Ω_0 and the odds that the case was exposed by Ω_1, and we have written K for the expression

$$\frac{1}{(1+\Omega_0)(1+\Omega_1)}$$

which occurs in all four probabilities. To obtain the hypergeometric likelihood we argue conditionally on the number of subjects exposed, N_1. It is clear from the figure that, when $N_1 = 2$, there is only one possible exposure configuration; the *conditional* probability of the observation is 1 and there is no contribution to the log likelihood. Similarly, there is no

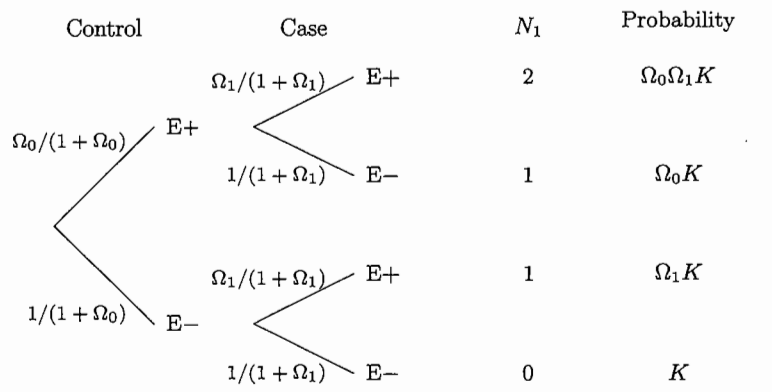

Fig. 19.2. Probabilities for a case-control set.

contribution to the log likelihood from sets in which $N_1 = 0$. These configurations correspond to the concordant sets which were also ignored in our previous analysis. However, when $N_1 = 1$ the exposure configuration could be either the second or third. These are the possible configurations of discordant sets. The observed split of discordant sets between the second and third configurations determines the log likelihood.

The conditional probabilities that a discordant set is of the third type (case exposed, control unexposed) and the second type (case unexposed, control exposed) are

$$\frac{\Omega_1 K}{\Omega_0 K + \Omega_1 K} \quad \text{and} \quad \frac{\Omega_0 K}{\Omega_0 K + \Omega_1 K}$$

respectively, and the conditional odds that the case was exposed is the ratio of these, Ω_1/Ω_0. This is the odds ratio parameter θ, assumed in our model to be constant for all the case control sets. The conditional argument therefore leads to a Bernoulli log likelihood based on splits of discordant sets into those in which the case is exposed and those in which the case is unexposed, the odds for such splits being θ. In our data, such sets split 15:7 and the log likelihood is

$$15\log(\theta) - 22\log(1+\theta).$$

Exercise 19.5. Calculate the most likely value of θ, a 90% confidence interval and the score test for the null hypothesis $\theta = 1$. These results of this exercise should agree precisely with those obtained using the Mantel–Haenszel method.

Table 19.2. Screening history in breast cancer deaths and matched controls

Status of the case	Number of controls screened			
	0	1	2	3
Screened	1	4	3	1
Unscreened	11	10	12	4

19.3 Several controls per case

The arguments outlined above may be extended to the situation in which there are several controls for each case. As before, we start with the Mantel–Haenszel approach.

Table 19.2 shows the results of a case-control study of breast cancer screening. Cases are deaths from breast cancer and each case is matched with three control women.† The exposure of interest is attendance for breast cancer screening. If screening is effective in prolonging life, screened women should have lower mortality rates and the odds ratio estimate from the case-control study should be less than 1. Note that as in Table 19.1, the table counts case-control sets and not women.

This study illustrates one of the reasons for matching discussed in Chapter 18. Women who die from breast cancer usually do so some years after initial diagnosis and during the period between diagnosis and death they would not be screened. Thus, controls would have a greater opportunity to be screened than cases. This difficulty was overcome by determining the relevant *exposure window*, the screening history of the controls was assessed over the period up to the time of diagnosis of the case, so that the screening histories of cases and controls are comparable. It was only possible to deal with this problem in this way because the study matched controls to individual cases.

Table 19.2 demonstrates the usual way such data are presented. However, it is very difficult to perceive any pattern — even as to whether or not screening appears to be a protective. To understand the analysis, we shall start by reordering the data as a tree. Fig. 19.3 illustrates the possible exposure configurations. The first three branches represent the exposure status of the three controls, the upper branch representing exposed (E+) and the lower unexposed (E−). Because we do not wish to differentiate between individual controls, this section of the tree may be abbreviated. For the first two controls, we do not need to differentiate between the configurations (E+, E−) and (E−, E+). These are simply grouped together as having 1 control exposed and we write the figure 2 at this point to remind us that branches emanating from this point are *double* branches. Similarly, after consideration of the third control we group together the 3 configurations with 2 exposed controls and the 3 configurations with 1 exposed control. The final branching represents the exposure status of the case.

†From Collette, H.J.A. *et al.* (1984) *The Lancet*, June 2, 1984, 1224–1226.

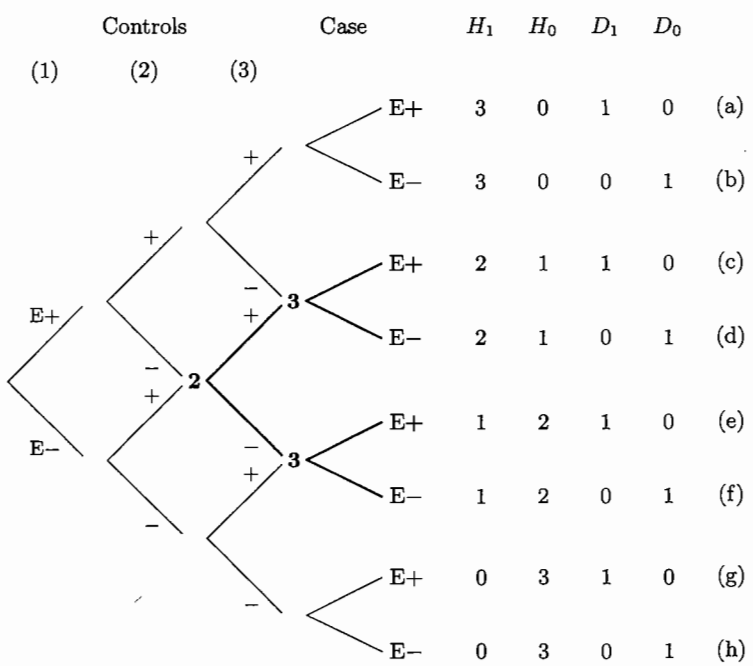

Fig. 19.3. Exposure configurations for 1:3 sets.

Exercise 19.6. In the screening data, how frequently do each of the eight types of exposure configuration occur?

We shall first analyse these data by the Mantel–Haenszel method. In the next section, we shall discuss the likelihood approach and show how it suggests a more useful arrangement of the table.

Exercise 19.7. Tabulate the values of Q, R, U, and V for these eight tables and hence calculate the Mantel–Haenszel significance test, odds ratio estimate and an approximate 90% confidence interval.

This analysis shows that the study finds a substantial and statistically significant reduction in mortality as a result of breast cancer screening.

19.4 The likelihood ⋆

The analysis of these data by use of the hypergeometric likelihood method is also quite straightforward. As before we argue conditionally upon the margins. Fig. 19.4 shows the total number of *subjects* exposed, N_1, and the

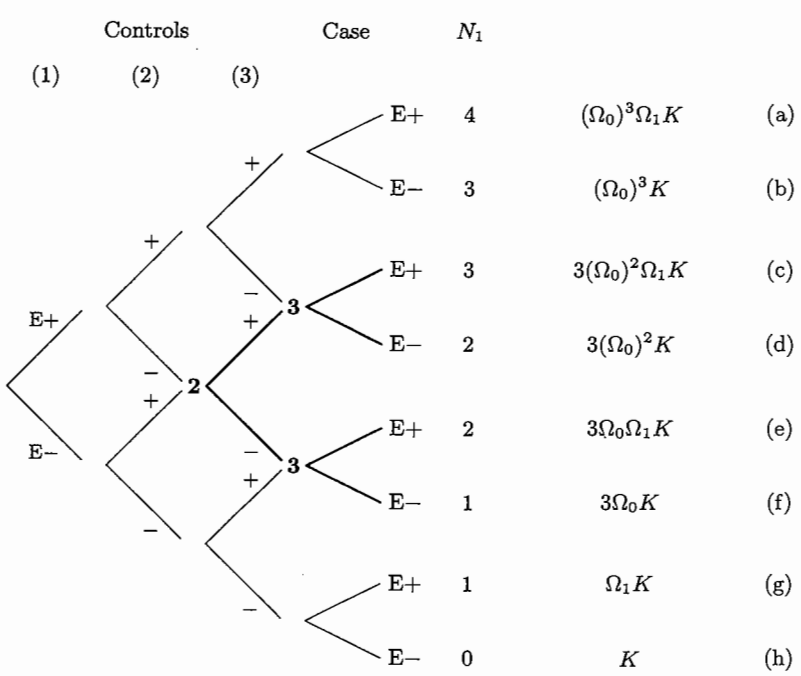

Fig. 19.4. Probabilities for 1:3 sets.

probability of each configuration, again writing K for the common factor, in this case

$$K = \frac{1}{(1+\Omega_0)^3(1+\Omega_1)}.$$

Note that the probabilities for configurations (c) to (f) are multiplied by 3 because each of these represents three paths in the complete tree. Now there are 5 possible values for the total number of subjects exposed. Again there are two *concordant* configurations in which the number of subjects exposed uniquely determines the configuration. $N_1 = 4$ ensures configuration (a) and $N_1 = 0$ ensures configuration (h). These make no contribution to the log likelihood. Each of the other three values of N_1 allows for two possible configurations, one in which the case is exposed and the other in which the case is unexposed. It is the splits of the observed data between these that yield the likelihood.

If the total number of exposed subjects in the set, N_1, is fixed at 3, then the exposure configuration must be either (b) or (c) and the conditional

Table 19.3. Splits of case-control sets

N_1	Split	Odds	Observed
3	(c):(b)	3θ	3:4
2	(e):(d)	θ	4:12
1	(g):(f)	$\theta/3$	1:10

odds for the split (c):(b) is

$$\frac{3(\Omega_0)^2\Omega_1 K}{(\Omega_0)^3 K} = \frac{3\Omega_1}{\Omega_0} = 3\theta.$$

Similarly, $N_1 = 2$ implies (d) or (e) and $N_1 = 0$ implies (f) or (g). The odds predicted by the model for these splits are set out in Table 19.3, together with the observed frequencies. By eye we can see that a value of θ of about 0.3 predicts the observed splits very well indeed. More formally, the log likelihood is

$$\begin{aligned} & 1\log\left(\frac{\theta}{3}\right) - 11\log\left(1+\frac{\theta}{3}\right) \\ + \; & 4\log(\theta) - 16\log(1+\theta) \\ + \; & 3\log(3\theta) - 7\log(1+3\theta). \end{aligned}$$

There is no simple expression for the maximum likelihood estimate and it is necessary to use a computer program to search for the maximum. This occurs at $\theta = 0.31$ ($\log(\theta) = -1.18$). The plot of the log likelihood ratio against $\log(\theta)$ is shown in Fig. 19.5. A Gaussian approximation with $S = 0.404$ fits quite closely.

The generalization of this argument to any number of controls per case may be carried out algebraically or by extending our tree. For sets of N_1 exposed subjects and N_0 unexposed subjects, the constant odds ratio model predicts that sets will split between those with an exposed case and those with an unexposed case with odds

$$N_1\theta/N_0.$$

A similar generalization is possible for several *cases* in each set. We will not give the details here, but computer software is readily available. Such analyses do not arise frequently in practice. An exception is family studies in which more than one sibling may be affected by a disease and unaffected siblings are used as controls.

In the examples discussed in this chapter, the Mantel–Haenszel and likelihood methods agree closely. The calculations for the former are rather easier, but the advantage of the likelihood approach lies in its greater generality and possibilities for extension. For example, when there are more

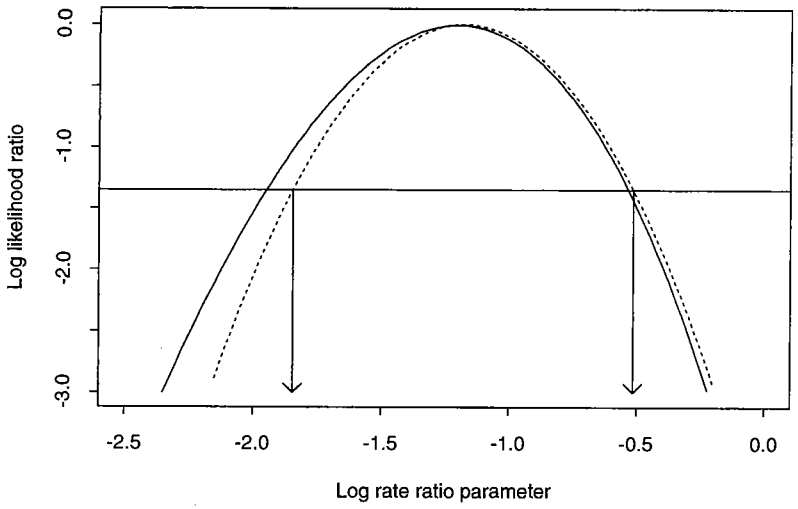

Fig. 19.5. Log likelihood ratio for $\log(\theta)$.

than two exposure categories, there is no simple method analogous to the Mantel–Haenszel approach. We shall defer discussion of such extensions to Part II of the book.

Solutions to the exercises

19.1 In the order in which the exposure configurations are listed in the figure, their frequencies are 26, 7, 15, and 37.

19.2 In the same order as listed,

Q	R	U	V
0	0	0	0
0	1/2	-1/2	1/4
1/2	0	1/2	1/4
0	0	0	0

Only the second and third configurations contribute to Q, R, U, and V.

19.3

$$\begin{aligned} Q &= 15 \times (1/2) \\ R &= 7 \times (1/2) \\ U &= 15 \times (1/2) - 7 \times (1/2) = 4 \end{aligned}$$

$$V = 15 \times (1/4) + 7 \times (1/4) = 5.5$$

The odds ratio estimate is $15/7 = 2.14$. This estimates the underlying rate ratio, so that the suggestion is that tonsillectomy doubles the rate of Hodgkin's disease. Using the expression

$$S = \sqrt{\frac{V}{QR}} = 0.4577,$$

the 90% error factor for the odds ratio is $\exp(1.645 \times 0.4577) = 2.12$. The 90% confidence limits are, therefore, $2.14/2.12 = 1.01$ (lower limit) and $2.14 \times 2.12 = 4.54$ (upper limit). Referring the value $(U)^2/V = 2.91$ to the chi-squared distribution gives $p \approx 0.09$.

19.4 If the matching is ignored, the following 2×2 table is obtained:

History:	Positive	Negative
Cases	41	44
Controls	33	52

The odds ratio in this table is $(41 \times 52)/(33 \times 44) = 1.47$, as compared to the value of 2.14 obtained by the correct analysis.

19.5 The most likely value is $15/7 = 2.14$. To calculate the approximate 90% interval using Gaussian approximation of the log likelihood for $\log(\theta)$ we use

$$S = \sqrt{\frac{1}{15} + \frac{1}{7}} = 0.4577,$$

the same as we obtained with the Mantel–Haenszel method. Under the null hypothesis, the probability for the split is 0.5 so that the expected number of sets with an exposed case is $22 \times 0.5 = 11$. The score and score variance are

$$\begin{aligned} U &= 15 - 11 = 4, \\ V &= 22 \times 0.5 \times 0.5 = 5.5. \end{aligned}$$

Again these are the values we obtained using the Mantel–Haenszel method.

19.6 In the order listed in the figure, the 8 exposure configurations have frequencies 1, 4, 3, 12, 4, 10, 1, 11.

19.7 The contributions to Q, R, U and V are shown below:

	Number of sets	Q	R	U	V
(a)	1	0	0	0	0
(b)	4	0	3/4	−3/4	9/48
(c)	3	1/4	0	1/4	9/48
(d)	12	0	2/4	−2/4	12/48
(e)	4	2/4	0	2/4	12/48
(f)	10	0	1/4	−1/4	9/48
(g)	1	3/4	0	3/4	9/48
(h)	11	0	0	0	0
Total		14/4	46/4	-32/4	354/48

Note that each contribution has to be multiplied by the number of times it occurred so that, for example, the total value of Q is

$$(3 \times 1/4) + (4 \times 2/4) + (1 \times 3/4) = 14/4.$$

The Mantel–Haenszel estimate of θ is $14/46 = 0.30$ and the chi-squared test is $(U)^2/V = 8.68$ ($p < 0.01$). An approximate error factor can be calculated from

$$\exp\left(1.645 \times \sqrt{\frac{V}{QR}}\right) = 2.02$$

so that the 90% confidence interval lies from $\theta = 0.15$ to $\theta = 0.60$.