17
Likelihoods for the odds ratio

The data from a simple case-control study (exposed and unexposed) can be arranged as a 2×2 table such as that set out in Table 17.1. We saw in Chapter 16 that there are two ways in which the probability model for a case-control study can be set up but that, for both models, the ratio of odds parameters are equal to the ratio of odds of failure in the study base.

17.1 The retrospective log likelihood

As in Chapter 16, we write Ω_0 for the odds of exposure among controls, and Ω_1 for the odds of exposure among cases. Our interest is in the odds ratio parameter $\theta = \Omega_1/\Omega_0$, so we change from the parameters Ω_0 and Ω_1 to the parameters Ω_0 and θ, and regard Ω_0 as a nuisance parameter. The total log likelihood is the sum of the log likelihood for Ω_0 based on the split of the H controls between exposed and unexposed, and the log likelihood for Ω_1 $(= \theta\Omega_0)$ based on the split of D cases,

$$H_1 \log(\Omega_0) - H \log(1 + \Omega_0) \quad + \quad D_1 \log(\theta\Omega_0) - D \log(1 + \theta\Omega_0).$$

To use this log likelihood for estimating of the odds ratio θ, we form a profile log likelihood by replacing Ω_0 by its most likely value for each value of θ. Unlike the profile log likelihood for the rate ratio in cohort studies, this curve cannot be expressed as a simple algebraic expression, but the results of section 13.4 and Appendix C can be used to derive a Gaussian approximation.

This derivation follows from the fact that the *log* odds ratio is the difference between two log odds parameters,

$$\log(\theta) = \log(\Omega_1) - \log(\Omega_0).$$

Table 17.1. Notation for the 2×2 table

Exposure	Cases	Controls	Total subjects
Exposed	D_1	H_1	$N_1 = D_1 + H_1$
Unexposed	D_0	H_0	$N_0 = D_0 + H_0$
Total	D	H	$N = D + H$

These are estimated from two independent bodies of data and have most likely values

$$M_1 = \log\left(\frac{D_1}{D_0}\right), \qquad M_0 = \log\left(\frac{H_1}{H_0}\right),$$

and standard deviations

$$S_1 = \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}, \qquad S_0 = \sqrt{\frac{1}{H_1} + \frac{1}{H_0}}.$$

It follows from general results given in section 13.4 and Appendix C that the most likely value of the log odds ratio is

$$\begin{aligned} M &= M_1 - M_0 \\ &= \log\left(\frac{D_1/D_0}{H_1/H_0}\right) \end{aligned}$$

and the standard deviation of the Gaussian approximation to the log likelihood is

$$\begin{aligned} S &= \sqrt{(S_1)^2 + (S_0)^2} \\ &= \sqrt{\frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{H_1} + \frac{1}{H_0}}. \end{aligned}$$

This can be used to calculate an error factor for the odds ratio and hence an approximate 90% confidence interval.

The expression for S only differs from that for the rate ratio in a cohort study by the addition of the two last terms. These are reciprocals of the counts of controls and represent the loss of precision incurred by carrying out a case-control study rather than a cohort study. Once the number of controls is substantially larger than the number of cases, this loss of precision becomes negligible. Hence the common assertion that there is little to be gained by drawing more than four or five times as many controls as cases.

Exercise 17.1. For the study of BCG vaccination and leprosy discussed in Chapter 16, calculate the expected result of the study using
(a) the same number of controls as cases;
(b) twice as many controls as cases; and
(c) five times as many control as cases.
Compare the corresponding values of S with that achieved by using the entire population as controls.

Carried out algebraically, these calculations lead to the general result that the ratio of the standard deviation of an estimate from a case-control study to the standard deviation from a cohort study yielding the same number

of cases is

$$\sqrt{1+(1/m)}$$

where m is the number of controls expressed as a multiple of the number of cases. When $m = 1$ this expression shows that the standard deviation is 1.41 times higher in a case-control study than in a cohort study. When $m = 5$ the factor reduces to 1.10 and when $m = 10$ this reduces only a little more to 1.05. The behaviour of this expression as m increases confirms the impression of the last exercise — that there is little gain in efficiency to be obtained by selecting more than five times as many controls as cases.

THE NULL HYPOTHESIS $\theta = 1$

We can calculate an approximate p-value for the null hypothesis using using any one of the three methods we have encountered earlier. The log likelihood ratio test is now based on the profile log likelihood. The Wald test is calculated by comparing the most likely value of the odds ratio with the null value, $\log(\theta) = 0$, by calculating

$$\left(\frac{M-0}{S}\right)^2.$$

Finally, the score test can be derived using the general relationships set out in Appendix C. At the null hypothesis the two odds parameters are equal and their most likely common value is N_1/N_0. The score, U, is found from the gradient of the profile log likelihood with respect to $\log(\Omega_1)$ at this point, which turns out to be

$$\begin{aligned} U &= D_1 - E_1 \\ &= -(D_0 - E_0), \end{aligned}$$

where

$$E_1 = D\frac{N_1}{N}, \qquad E_0 = D\frac{N_0}{N}$$

can be thought of as the expected numbers of exposed and unexposed cases under the null hypothesis. The score variance is obtained from the curvature of the profile log likelihood at the null value $\theta = 1$, which yields

$$V = \frac{DHN_0N_1}{(N)^3}.$$

As usual, an approximate p-value can be obtained by referring $(U)^2/V$ to the chi-squared distribution on one degree of freedom.

Table 17.2. Tonsillectomy and Hodgkins disease

Tonsillectomy	Cases		Controls		Total subjects	
Positive	90	(D_1)	165	(H_1)	255	(N_1)
Negative	84	(D_0)	307	(H_0)	391	(N_0)
Total	174	(D)	472	(H)	646	(N)

Exercise 17.2. Table 17.2 shows data from a study of the relationship between tonsillectomy and the incidence of Hodgkin's disease.* Calculate the maximum likelihood estimate of θ with a 90% confidence interval, and calculate a p-value for $\theta = 1$.

17.2 The prospective log likelihood

We now turn to the log likelihood we obtain using the prospective probability model. As in Chapter 16, we write ω_1 for the odds that an exposed subject is a case, ω_0 for the corresponding odds for an unexposed subject, and change to (ω_0, θ) where $\theta = \omega_1/\omega_2$. The log likelihood is again the sum of two Bernoulli log likelihood terms,

$$D_0\log(\omega_0) - N_0\log(1+\omega_0) \quad + \quad D_1\log(\theta\omega_0) - N_1\log(1+\theta\omega_0),$$

and the profile log likelihood is obtained by replacing ω_0 by its most likely value at each value of θ. As with the retrospective model, this does not lead to a simple algebraic expression, but the Gaussian approximation can easily be derived, since

$$\log(\theta) = \log(\omega_1) - \log(\omega_0)$$

and the log likelihoods for $\log(\omega_1)$ and $\log(\omega_0)$ are based on independent sets of data. The most likely values of ω_1 and ω_0 are

$$M_1 = \log\left(\frac{D_1}{H_1}\right), \qquad M_0 = \log\left(\frac{D_0}{H_0}\right),$$

and the corresponding standard deviations are

$$S_1 = \sqrt{\frac{1}{D_1}+\frac{1}{H_1}}, \qquad S_0 = \sqrt{\frac{1}{D_0}+\frac{1}{H_0}}.$$

As before, the most likely value of $\log(\theta)$ is

$$M \quad = \quad M_1 - M_0$$

*From Johnson, S.K. and Johnson, R.E. (1972) *New England Journal of Medicine*, **287**, 1122–1125.

$$= \log\left(\frac{D_1/H_1}{D_0/H_0}\right)$$

and the standard deviation of the Gaussian approximation to the log likelihood is

$$\begin{aligned} S &= \sqrt{(S_1)^2 + (S_0)^2} \\ &= \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}}. \end{aligned}$$

These results are exactly the same as we obtained using the retrospective argument. In the same way we can show that the log likelihood ratio and score tests are identical for the two approaches. Indeed, some further mathematics shows that the profile log likelihoods for the two arguments are identical. This continues to be the case for more complex patterns of exposure and, since the prospective approach is more convenient in these situations, it is to be preferred.

17.3 The hypergeometric likelihood

Both the probability models discussed above contain a nuisance parameter in addition to the parameter of interest, θ. Both lead to profile log likelihood for θ and depend on profile likelihood behaving in the same way as a true likelihood.

When there is sufficient data, the profile log likelihood does indeed behave in this way. However, profile likelihoods are obtained by estimating the nuisance parameters, and it is only safe to assume that they have the same properties as true likelihoods if the accuracy of that estimation increases as the total number of subjects increases. If the number of nuisance parameters increases with the number of subjects, this improved estimation is not achieved and profile likelihoods can be misleading. This happens in case-control studies if, as the total number of subjects increases, the study is divided into an increasing number of small strata in an attempt to deal with confounding. For either the prospective or the retrospective likelihood it is necessary to introduce a separate nuisance parameter for each stratum, so the number of parameters will increase with the number of subjects. The worst case is the individually matched case-control study in which the number of strata (and nuisance parameters) is equal to the number of case-control pairs. It turns out that the use of profile likelihood methods in this situation leads to wrong answers.

An alternative way of eliminating the nuisance parameter is a *conditional* approach based on a probability model in which *both* margins of the 2×2 table (Table 17.1) are fixed. The set of probabilities for all splits of subjects which maintain the same marginal totals is known as the *hypergeometric* distribution. For the table shown in Table 17.1, the probability

is

$$\frac{1}{\mathrm{K}(\theta)} \times \frac{(\theta)^{D_1}}{D_1!D_0!H_1!H_0!}$$

where $\mathrm{K}(\theta)$ is chosen so that the probabilities for all possible tables with the same margins add up to one:

$$\mathrm{K}(\theta) = \sum_{\text{Possible tables}} \frac{(\theta)^{D_1}}{D_1!D_0!H_1!H_0!}.$$

This distribution depends only on the parameter θ and can be used to calculate exact p-values and confidence intervals for the odds ratio as outlined in Chapter 12. The use of these methods is illustrated in section 17.4.

The likelihood based on this distribution is called the *hypergeometric likelihood*. Because of the function $\mathrm{K}(\theta)$, it is difficult to calculate except when the number of possible tables consistent with the margins is small. We shall consider an important special case in Chapter 19 and give a more general treatment of this likelihood in Chapter 29. For the present we note that the hypergeometric likelihood does lead to a simple score test for $\theta = 1$. The score is exactly the same as for the profile log likelihoods, that is

$$U = D_1 - E_1,$$

but the score variance can be shown to be

$$V = \frac{DHN_0N_1}{(N)^2(N-1)}.$$

This differs from the expression derived from the curvature of the *profile* log likelihood by the term $(N-1)$ in place of N in the denominator. The difference this makes to the value of the variance is usually negligible. The one situation where it does make a difference is in matched studies where the number of subjects in each stratum is very small. In the worst case of the 1:1 individually matched study, $N = 2$ in every stratum and the profile likelihood argument wrongly estimates the score variance by a factor of two. We shall, therefore, return to the hypergeometric likelihood when discussing the analysis of individually matched case-control studies in Chapter 19.

17.4 Exact methods ⋆

The use of the hypergeometric distribution for exact tests and confidence intervals follows exactly the same principles as set out in Chapter 12. This is illustrated in this section using some data drawn from a case-control study set up to investigate an excess of childhood leukaemia cases in the

Table 17.3. Paternal radiation exposure in leukaemia cases and controls

Paternal exposure	Leukaemia cases	Local controls	Total
$\geq$ 100 mSv (Exposed)	3	1	4
$<$ 100 mSv (Unexposed)	1	19	20
Total	4	20	24

Table 17.4. Hypergeometric log likelihood ratios and probabilities

	LLR	Hypergeometric probability		
D_1	$(\theta = 1)$	$(\theta = 1)$	$(\theta = 2.440)$	$(\theta = 1534.1)$
0	-0.785	0.455957	0.202245	
1	-0.105	0.429136	0.464450	0.000001
2	-1.451	0.107284	0.283314	0.000460
3	-4.252	0.007529	0.048511	0.049540
4	-9.271	0.000094	0.001480	0.949998
Total		1.0	1.0	1.0

vicinity of a nuclear reprocessing plant (see Exercise 11.8). The data set out in Table 17.3 concern occupational radiation exposure in fathers of 4 cases and fathers of 20 local controls.†

There are five possible tables with the same margins as Table 17.3, with values of D_1 (the number of exposed cases) ranging from zero to four. The hypergeometric distribution gives the conditional probability for each table as a function of the odds ratio parameter, θ, and the log likelihood for any value of θ is calculated by taking the log of the probability of the *observed* outcome $D_1 = 3$. The most likely value of θ is 37.345‡ and the log likelihood ratio which compares this with the null value $(\theta = 1)$ is -4.252. Table 17.4 shows, in the column headed LLR, similar log likelihood ratio comparisons for each of the five possible tables and, in the next column, the conditional probabilities of these tables when the null hypothesis is true. The p-value is the sum of probabilities of the observed table and of all tables which are in greater conflict with the null value. In this case $p = 0.007529 + 0.000094 = 0.007623$. The one-sided and two-sided p-values are identical in this case. This way of calculating the p-value for a 2×2 table is called *Fisher's exact test.*

Similar ideas are used to calculate 'exact' confidence intervals. To find

†From Gardner, M.J. *et al.* (1990) *British Medical Journal*, **300**, 423–429.

‡Note that this is not the same value as that obtained with the profile likelihood which is $(3 \times 19)/(1 \times 1) = 57$.

the limits of the 90% interval we search for values of θ which give one-sided p-values of 0.05. These values are 2.440 (lower limit) and 1534.1 (upper limit) and the corresponding hypergeometric distributions are shown in the last two columns of Table 17.4. At $\theta = 2.440$ the one-sided p-value is $0.048511 + 0.001480 = 0.04991$ and at $\theta = 1534.1$ the one-sided p-value is $0.000001 + 0.000460 + 0.049540 = 0.050001$. Values of θ outside the range from 2.440 to 1534.1 would have smaller p-values than 0.05 and the frequentist theory would therefore suggest that we should pronounce ourselves 90% confident that θ lies within this range. As we have seen in Chapter 12, this is a very technical use of the word confident and no epidemiologist would really believe that θ could really take such large values. The extreme finding is obtained, at least to some extent, because the radiation level chosen here to divide exposed and unexposed groups was chosen *after* seeing the data.

Solutions to the exercises

17.1 The following shows the expected results of the three studies. These have been calculated by splitting the controls between scar present and scar absent categories in the proportions 46 028/80 622 and 34 594/80 622 respectively.

			Expected controls		
BCG scar	Cases	Population	(a)	(b)	(c)
Present	101	46 028	148	296	740
Absent	159	34 594	112	224	560
Total	260	80 622	260	520	1300

The standard deviations for the log odds ratio estimate are worked out using the formula $S = \sqrt{1/D_0 + 1/D_1 + 1/H_0 + 1/H_1}$ and are 0.179, 0.155, and 0.139 respectively. The standard deviation using the full data is 0.127. The gain in precision with increasing numbers of controls clearly follows a law of diminishing returns.

17.2 The maximum likelihood estimate of θ is the observed odds ratio:

$$\frac{90/84}{165/307} = 1.99.$$

and

$$S = \sqrt{\frac{1}{84} + \frac{1}{90} + \frac{1}{307} + \frac{1}{165}} = 0.180.$$

For calculating 90% confidence limits, the error factor is $\exp(1.645 \times 0.180) = 1.34$. The limits are therefore $1.99/1.34 = 1.48$ (lower limit) and $1.99 \times 1.34 = 2.67$ (upper limit).

The expected number of exposed cases is given by

$$E_1 = 174 \times \frac{255}{646} = 68.68$$

so that the score, U, is $(90 - 68.68) = 21.32$. The score variance is

$$\frac{174 \times 472 \times 255 \times 391}{(646)^3} = 30.37.$$

The score test is $(21.32)^2/30.37 = 14.97$, $(p < 0.001)$.