13
Likelihoods for the rate ratio

In previous chapters we have introduced the main ideas of probability models in epidemiology and discussed the use of likelihood to provide an estimate, confidence interval or p-value for the parameter of a probability model. Although we have used the joint log likelihood for several parameters our discussion of confidence intervals and p-values has been based on probability models with only a single parameter. We now consider probability models with two or more parameters.

13.1 Comparing rates using the rate ratio

A simple and important problem which involves two parameters is the comparison of two rates, for example for a cohort which was exposed to some environmental factor and an unexposed cohort. The probability model which underlies such a comparison has parameters corresponding to the rates of failure in the two cohorts. We shall use a subscript notation to denote exposure groups and write λ_1 for the rate parameter conditional on exposure, and λ_0 for the rate parameter conditional on non-exposure.

Table 13.1 shows a preliminary tabulation of some data which will be analysed in detail in this and the following chapter.* The data relate subsequent incidence of ischaemic heart disease (IHD) to dietary energy intake. The study cohort consisted of 337 men whose energy intake was assessed by a seven-day weighed dietary survey. The subsequent follow-up was for an average of 13.7 years and yielded 45 new cases of IHD. The table divides this cohort into an exposed group consisting of men whose energy intake was less than 2750 kcals per day, the remaining men being regarded as unexposed. Although it might seem odd to denote the low energy intake group as exposed, this is because low energy intake is a surrogate measure for physical inactivity. Table 13.1 also introduces some algebraic notation: D_0, D_1 for the number of disease events observed in the unexposed and exposed cohorts respectively, and Y_0, Y_1 for the corresponding person-years observation.

*Unpublished data. The study is described by Morris, J.N. *et al.* (1977) *British Medical Journal*, 19 November 1977, **2**, 1307–1314.

Table 13.1. Incidence of ischaemic heart disease by energy intake

	Energy intake	
	< 2750 kcals (exposed)	≥ 2750 kcals (unexposed)
Person years	1857.5 (Y_1)	2768.9 (Y_0)
New cases	28 (D_1)	17 (D_0)
Estimated rate	15.1	6.1
90% interval	$(11.1 \to 20.6)$	$(4.1 \to 9.1)$

The data from the unexposed group leads to

$$D_0 \log(\lambda_0) - \lambda_0 Y_0 = 17 \log(\lambda_0) - 2\,768.9\lambda_0$$

as the log likelihood for λ_0. The most likely value of λ_0 is the observed incidence rate, $17/2768.9 = 6.1$ per 1000 person-years. The fact that this estimate is based on only 17 observed cases is reflected in the rather wide 90% confidence interval for λ_0 stretching from 4.1 to 9.1 per 1000 person-years. Similarly, the data from the exposed group leads to

$$D_1 \log(\lambda_1) - \lambda_1 Y_1 = 28 \log(\lambda_1) - 1857.5\lambda_1$$

as the log likelihood for λ_1. The most likely value of λ_1 is $28/1857.5 = 15.1$ per 1000 person-years, and the 90% confidence interval stretches from 11.1 to 20.6 per 1000 person-years. The two groups provide independent sets of data, so that the two log likelihoods are added to yield the joint log likelihood

$$17 \log(\lambda_0) - 2\,768.9\lambda_0 + 28 \log(\lambda_1) - 1\,857.5\lambda_1.$$

This is the likelihood for any specified pair of values for the two parameters λ_0 and λ_1. Its maximum value is achieved when these parameters take values equal to the corresponding observed rates — 6.1 and 15.1 per 1000 person-years respectively.

The 90% confidence intervals for the two rates do not overlap and it might seem that the data support the proposition that the two rates are different. In general, however, the degree of overlap of confidence intervals is a poor criterion for comparing rates. If the interval in the high intake group had stretched from, say, 3.0 to 12.0 then it could be argued that, since values of the rate parameter in the range from 11.1 to 12.0 are included in both intervals, the data do not support the idea that the rates are different. The flaw in this argument is that this range is at the extreme of both ranges; the support for the proposition that the rates are similar requires two rather poorly supported propositions to hold simultaneously.

The way to approach such problems is to reparametrize the model in such a way that one of the new parameters makes a comparison. The usual comparison parameter for two rates is the *rate ratio*, which we shall denote by the Greek letter θ. Since $\theta = \lambda_1/\lambda_0$, the rate in the exposed cohort may be written as $\theta\lambda_0$ instead of λ_1 and our model can be written in terms of the parameters (θ, λ_0) instead of (λ_1, λ_0).

The log likelihood for λ_0 and λ_1 in terms of D_0, D_1, Y_0, Y_1 is

$$D_0 \log(\lambda_0) - \lambda_0 Y_0 + D_1 \log(\lambda_1) - \lambda_1 Y_1.$$

To express the log likelihood in terms of the new parameter system, we substitute $\theta\lambda_0$ for λ_1, to get

$$D_0 \log(\lambda_0) - \lambda_0 Y_0 + D_1 \log(\theta\lambda_0) - \theta\lambda_0 Y_1,$$

which reduces to

$$D \log(\lambda_0) + D_1 \log(\theta) - \lambda_0 Y_0 - \theta\lambda_0 Y_1,$$

where $D = D_0 + D_1$ is the total number of observed disease events. For the example in Table 13.1, the log likelihood is

$$45 \log(\lambda_0) + 28 \log(\theta) - 2768.9\lambda_0 - 1857.5\theta\lambda_0$$

The purpose of this choice of new parameters for the model is to concentrate the comparison of the rates into the parameter θ, but unfortunately, the log likelihood for these new parameters cannot be divided into a sum of separate parts, one for each parameter. The appearance of the term $1857.5\theta\lambda_0$ means that the shape of the log likelihood with respect to θ depends on the value of λ_0, and this is unknown. When assessing the support for different values of θ, not knowing λ_0 is somewhat of a problem and in this context λ_0 is called a *nuisance parameter*.

There are two ways of dealing with a nuisance parameter when constructing a likelihood for the parameter of interest. These will be described in the next two sections.

13.2   Profile likelihood

The obvious way to deal with a nuisance parameter is to *estimate* its value. For each value of the rate ratio θ, the value of λ_0 which maximizes the likelihood can be determined and substituted into the joint log likelihood. The resulting maximized log likelihood can then used as a measure of support for this value of θ.

This idea is illustrated in Fig. 13.1. The top graph shows the log likelihood ratio for $\log(\lambda_0)$ and $\log(\theta)$ as a contour map. The contour lines,

corresponding to parameter values which have equal log likelihood, are approximately elliptical (this has been aided by the choice of log scales for both parameters, so that they are not bounded). The contours shown correspond to log likelihood ratios of -1, -2, -3, -4, and -5 relative to the maximum value.

The vertical arrows denote specified values of $\log(\theta)$ for which we require to measure the support. For each fixed value of $\log(\theta)$, we find the value of $\log(\lambda_0)$ which maximizes the log likelihood and plot this maximized log likelihood on the lower graph. This is then used to measure the relative support lent by the data to different values of $\log(\theta)$. By analogy with physical maps, this curve is called a *profile* log likelihood. A profile log likelihood is not a true log likelihood since it cannot be directly obtained by taking the log of the probability of the data. However, in most situations it behaves in exactly the same way as a log likelihood. It can be seen from Fig. 13.1 that the value of θ which gives the largest value of the profile log likelihood is also the value corresponding to the maximum of the total log likelihood. The curvature of the profile log likelihood at this maximum point can be used to calculate approximate confidence intervals and Wald tests, and score tests for null values of θ can be carried out using the gradient and curvature of the profile log likelihood at the null value. Similarly, a log likelihood ratio test can be carried out by calculating minus twice the profile log likelihood ratio at the null value of θ.

In the case of the the rate ratio, this process is simplified since the derivation of the profile log likelihood can be carried out algebraically, leading to a mathematical equation for the curve. The value of λ_0 which maximizes the log likelihood for any given value of θ may be shown to be

$$\frac{D}{Y_0 + \theta Y_1}$$

and substituting this for λ_0 in the log likelihood expression gives the profile log likelihood:

$$D_1 \log(\theta) - D \log(Y_0 + \theta Y_1) + D \log(D) - D.$$

Since the last two terms do not depend upon θ, they are irrelevant and may be omitted. We are also at liberty to *add* terms which do not involve θ, and addition of

$$D_1 \log(Y_1) + D_0 \log(Y_0)$$

yields, after some rearrangement, the expression:

$$D_1 \log\left(\frac{\theta Y_1}{Y_0}\right) - D \log\left(1 + \frac{\theta Y_1}{Y_0}\right).$$
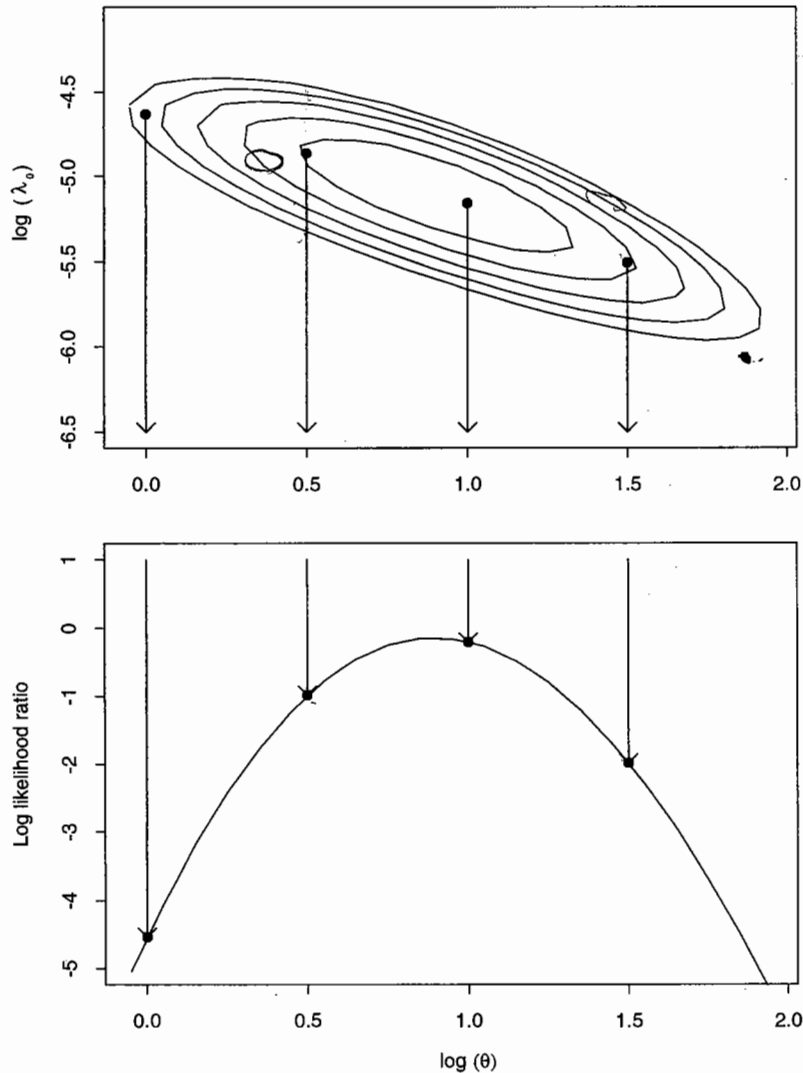
Fig. 13.1.   Log likelihood surface for θ and λ (above) and profile log likelihood for θ (below).

This is exactly the same as a Bernoulli likelihood for the odds parameter

$$\Omega = \frac{\theta Y_1}{Y_0}$$

based on a split of D cases as D_1 exposed and $D_0 = D - D_1$ unexposed. It follows that estimation of θ using the profile log likelihood is equivalent to estimating the odds, Ω, in the binary model; the two estimates differ only by the known multiplier, Y_1/Y_0.

From the Bernoulli likelihood, the most likely value of Ω is D_1/D_0 and the standard deviation of $\log(\Omega)$ is

$$\sqrt{\frac{1}{D_0} + \frac{1}{D_1}}.$$

It follows that the most likely value of θ is

$$\frac{D_1/D_0}{Y_1/Y_0} = \frac{D_1/Y_1}{D_0/Y_0}$$

which is the ratio of the most likely values of the two rates and since $\log(\theta)$ differs from $\log(\Omega)$ only by a known constant, the shape of the log likelihoods are identical, and the standard deviation of $\log(\theta)$ is also

$$\sqrt{\frac{1}{D_0} + \frac{1}{D_1}}.$$

Exercise 13.1. Calculate the maximum likelihood estimate of the rate ratio for the data of Table 13.1 and give 90% confidence limits.

For the calculation of p-values, the null hypothesis generally of interest is that the two rates are equal, so that $\theta_\oslash = 1$ and $\Omega_\oslash = Y_1/Y_0$. In terms of the corresponding risk parameter the null hypothesis is that

$$\pi_\oslash = \frac{\Omega_\oslash}{1 + \Omega_\oslash} = \frac{Y_1}{Y_0 + Y_1}.$$

The score is

$$U = D_1 - D\pi_\oslash,$$

which can be written as

$$U = D_1 - E_1$$

where $E_1 = D\pi_\oslash$ is the expected number of exposed cases under the null hypothesis. The score variance is

$$V = D\pi_\oslash(1 - \pi_\oslash).$$

Exercise 13.2. Test the significance of the effect of low energy intake in the data of Table 13.1.

13.3 Conditional likelihood

The approach outlined above starts from the question: what is the probability that, during follow-up, D_0 events occur in the unexposed cohort and D_1 in the exposed cohort? The resulting likelihood involves not only the rate ratio θ (the parameter of interest), but also a nuisance parameter, λ_0. Replacing the unknown nuisance parameter by its most likely value leads to the profile log likelihood for θ. This argument is appealing in that it closely follows the way in which cohort studies are designed and executed — we decide in advance upon the cohort to be followed and the duration of follow-up and wait to see how many disease events occur in different subgroups. However, it is not essential that the likelihood argument should correspond so closely with the study design. In particular, if some aspect of the result contains little or no information about the parameter of interest, then we are free to treat it as if it were fixed by the study design. The aim of such an argument, which is called a *conditional argument*, is to obtain a new probability model for the data which does not involve the nuisance parameter.

In this case the total number of cases tells us nothing about the effect of exposure, which depends on the split among cases between exposed and not exposed. We therefore take the total number of events as fixed, corresponding to a study in which the follow-up continues for just long enough for D events to be observed. The analysis of the study then concentrates on the split of cases between the exposed and unexposed sections of the cohort, and starts from the question: given that D failures occurred, what is the probability that D_0 of them occurred in the unexposed group and D_1 in the exposed group?

The split of the D failures between exposed and unexposed groups may be described using the binary probability model. This is illustrated in Fig. 13.2. The left-hand tree shows the observed split of the failures and the right-hand tree shows the expected split of cases. If Y_1 and Y_0 can be regarded as fixed, the odds that a case was exposed is

$$\Omega = \frac{\lambda_1 Y_1}{\lambda_0 Y_0} = \frac{\theta Y_1}{Y_0},$$

and the log likelihood for θ is

$$D_1 \log\left(\frac{\theta Y_1}{Y_0}\right) - D \log\left(1 + \frac{\theta Y_1}{Y_0}\right).$$

Thus regarding the number of cases as fixed leads directly to a *conditional* log likelihood which depends only upon θ. The log likelihood is conditional
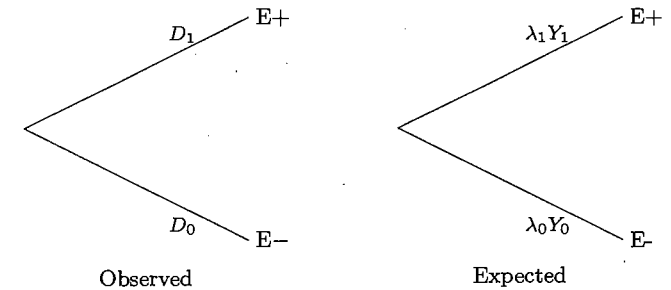

Fig. 13.2. The conditional argument.

in the sense that it takes as fixed an aspect of the data (the total number of events, D) that was, in reality, an unpredictable outcome of the study. In this case the profile and conditional likelihood approaches have led to the same log likelihood and, therefore, to identical estimates and confidence intervals, but in general this will not be the case.

The conditional approach always yields a true log likelihood, being based upon a probability (albeit a *conditional* probability) of observed data. Also, because this probability depends only on the parameter of interest, it can be used to calculate exact p-values and confidence intervals. In our current example, the probabilities for different splits of cases between exposed and unexposed groups, given θ, can be obtained from the binomial distribution. However, the conditional approach is not an automatic method, but relies on our ingenuity in recognizing a suitable conditional argument. Such arguments are not always possible. For example, it has not proved possible to find an argument which leads to a conditional likelihood for the *rate difference*, $\lambda_1 - \lambda_0$.

In contrast, the profile method has the considerable virtue that it can *always* be employed. Even if it is impossible to use an algebraic method to obtain an explicit formula for the profile log likelihood curve, the derivation of the curve numerically by the procedure illustrated in Fig. 13.1 can always be carried out by computer. The difficulty with this approach is that the profile curve is not necessarily a true log likelihood. However, in most situations it does approximately possess the properties of a true log likelihood. These properties can safely be assumed when the number of nuisance parameters is small in comparison with the total quantity of data.

We should note that our current use the conditional approach requires $\lambda_0 Y_0$ and $\lambda_1 Y_1$, the expected numbers of cases in the two groups, to be constants not influenced by the study outcome. Although this is approximately true for the rare events usually studied by epidemiologists (see section 6.3), it may not be an acceptable argument when the probabilities of failure are high. In these cases, the likelihood derived in this chapter

can only be regarded as a profile likelihood and exact tests and confidence intervals are not available.

★ 13.4 Approximating profile log likelihoods

For the rate ratio it is possible to derive a *mathematical expression* for the profile log likelihood and hence find a Gaussian approximation from which approximate p-values and confidence intervals can be calculated in the usual way. This is not possible in general. The profile likelihood can always be computed by going through the steps indicated in Fig. 13.1, but the resulting curve usually cannot be represented by a simple algebraic expression. Fortunately some simple rules, derived from calculus, allow us to calculate Gaussian approximations to such profile log likelihoods, and hence algebraic expressions for M, S, U, and V, which we can go on to use in the usual way. These rules and their derivation are explained in Appendix C. Here we briefly summarize the most important rules.

An important general problem is the estimation of the difference between two parameters β_0 and β_1 when these are estimated from two independent bodies of data. If the log likelihood for β_0 has a Gaussian approximation defined by the most likely value M_0 and standard deviation S_0 and the approximation to the log likelihood for β_1 is defined by M_1 and S_1, then the Gaussian approximation of the log likelihood for $\beta_1 - \beta_0$ has

$$
\begin{aligned}
M &= M_1 - M_0, \\
S &= \sqrt{(S_1)^2 + (S_0)^2}.
\end{aligned}
$$

The rate ratio is a special case of this more general problem since its logarithm may be written

$$
\log\left(\frac{\lambda_1}{\lambda_0}\right) = \log(\lambda_1) - \log(\lambda_0)
$$

and in Appendix C it is shown that these rules lead to the same Gaussian log likelihood approximation as we obtained earlier. Here we use them to approximate the profile log likelihood for the rate difference. The most likely value is the difference between the most likely values of the rates,

$$
M = \frac{D_1}{Y_1} - \frac{D_0}{Y_0},
$$

and, from Chapter 9, $S_1 = \sqrt{D_1}/Y_1$ and $S_0 = \sqrt{D_0}/Y_0$ so the value of S for the rate difference is

$$
\sqrt{\frac{D_1}{(Y_1)^2} + \frac{D_0}{(Y_0)^2}}.
$$

Exercise 13.3. Calculate an approximate 90% confidence interval for the difference between the rates using the data of Table 13.1.

A still more general problem concerns a weighted sum of parameters, of the form

$$
W_1\beta_1 + W_2\beta_2 + W_3\beta_3 + \cdots \qquad ,
$$

each β parameter again being estimated from independent bodies of data. The Gaussian approximation to the profile log likelihood for the weighted sum has

$$
\begin{aligned}
M &= W_1M_1 + W_2M_2 + W_3M_3 + \cdots \\
S &= \sqrt{(W_1S_1)^2 + (W_2S_2)^2 + (W_3S_3)^2 + \cdots} \qquad ,
\end{aligned}
$$

where $M_1, S_1, \ldots$ etc. are the most likely values and standard deviations for $\beta_1, \ldots$ etc.. An example is the profile log likelihood for the cumulative failure rate. In Chapter 5 we defined the cumulative rate by

$$
\lambda^1 T^1 + \lambda^2 T^2 + \cdots
$$

where $\lambda^1, \lambda^2, \ldots$ are probability rates operating for time periods $T^1, T^2, \ldots$. The cumulative rate is, therefore, a weighted sum of the form discussed in this section.

Exercise 13.4. Using the Gaussian approximation given in Chapter 9 for the log likelihoods for rate parameters, derive an expression for the Gaussian approximation to the profile log likelihood for the cumulative rate.

Solutions to the exercises

13.1 The most likely value of θ is

$$
\frac{D_1/Y_1}{D_0/Y_0} = \frac{28/1857.5}{17/2768.9} = 2.48.
$$

The standard deviation of the estimate of $\log(\theta)$, is

$$
S = \sqrt{1/28 + 1/17} = 0.3075,
$$

so that the 90% error factor for θ is

$$
\exp(1.645 \times 0.3075) = 1.66.
$$

The 90% confidence limits for the rate ratio are $2.48/1.66 = 1.49$ (lower limit) and $2.48 \times 1.66 = 4.12$ (upper limit).

13.2   The observed number of events in the low energy intake group is 28. There were 45 events in total and, under the null hypothesis, the probability of having been exposed is $\pi_0 = 1857.5/4626.4 = 0.402$. The score is

$$U = 28 - 45 \times 0.402 = 9.93,$$

and the score variance is

$$V = 45 \times 0.402 \times (1 - 0.402) = 10.81.$$

The score test is $(U)^2/V = 9.12$, giving $p \approx 0.003$.

13.3

$$M = \frac{28}{1857.5} - \frac{17}{2768.9} = 0.00893 \ (8.93 \text{ per } 1000 \text{ person-years}).$$

$$S = \sqrt{\frac{28}{(1857.5)^2} + \frac{17}{(2768.9)^2}} = 0.00321 \ (3.21 \text{ per } 1000 \text{ person-years}).$$

The 90% confidence interval is

$$M \pm 1.645S = 3.65 \text{ to } 14.2 \text{ per } 1000 \text{ person-years}.$$

13.4   The log likelihood for λ^1 is approximated by a Gaussian curve with

$$M^1 = \frac{D^1}{Y^1}, \qquad S^1 = \frac{\sqrt{D^1}}{Y^1}.$$

Similarly for $\lambda^2, \lambda^3, \ldots$ etc. The weights are the durations of observation, $T^1, T^2, \ldots$, so that the profile log likelihood for the cumulative rate has its maximum at

$$M = \frac{D^1}{Y^1}T^1 + \frac{D^2}{Y^2}T^2 + \cdots$$

and the standard deviation of the Gaussian approximation is

$$S = \sqrt{D^1 \left(\frac{T_1}{Y^1}\right)^2 + D^2 \left(\frac{T^2}{Y^2}\right)^2 + \cdots}.$$

Note that, as we narrow the time bands to clicks, the ratio T/Y approaches $1/N$, where N is the number of subjects under observation during the click. In these circumstances, M is the Aalen–Nelson estimate of the cumulative rate and S may be used to calculate an approximate confidence interval.

14
Confounding and standardization

14.1   Confounding

Epidemiological studies generally involve comparing the outcome over a period of time for groups of subjects experiencing different levels of exposure. Such studies are usually not controlled experiments but 'experiments of nature' of which the epidemiologist is a passive observer. In such investigations, there is always the possibility that an important influence on the outcome, which would have been fixed in a controlled experiment, differs systematically between the comparison groups. It is then possible that part of an apparent effect of exposure is due to these differences, and the comparison of the exposure groups is said to be *confounded*. Statistical approaches to dealing with the problem of confounding aim to correct, during analysis, for such deficiencies in the design of experiments of nature.

A particularly important potential confounding variable (or *confounder* in many epidemiological studies is the age of subjects. We shall consider an example in which subjects in a follow-up study are classified according to whether their age at the start of follow-up was less than 55 years or 55 years or more. Suppose that the breakdown between the two age groups is $0.8 : 0.2$ and that the conditional probability of failure is 0.1 in the first age group and 0.3 in the second. When age is ignored the overall or *marginal* probability of failure is

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14.$$

Now suppose that the age distribution differs between the two exposure groups, being $0.8 : 0.2$ in the not exposed group but $0.4 : 0.6$ in the exposed group (see Fig. 14.1). The marginal probability of failure for the unexposed group is still

$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14,$$

but for the exposed group it is now

$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22.$$

The marginal probabilities of failure now suggest an apparent effect of

13 Likelihoods for the rate ratio

13.1 Two Rates and a Rate Ratio

• The article referred to in the footnote had Clayton as co-author.[1]

• "This (the one shown) is the likelihood for any specified pair of values for the two parameters λ_0 and λ_1.' [...] Its maximum value is achieved when these parameters take values equal to the corresponding observed rates ($\widehat{\lambda_0}$ and $\widehat{\lambda_1}$)"

This begs the question as to whether the MLE of $\theta = \lambda_1/\lambda_1$ is $\widehat{\lambda_1}/\widehat{\lambda_0}$. It is in this case, but is it always the case?

• "The 90% confidence intervals for the two rates do not overlap and it might seem that the data support the proposition that the two rates are different. In general, however, the degree of overlap of confidence intervals is a poor criterion for comparing rates. If the interval in the high intake group had stretched from, say, 3.0 to 12.0 then it could be argued that, since values of the rate parameter in the range from 11.1 to 12.0 are included in both intervals, the data do not support the idea that the rates are different. The flaw in this argument is that this range is at the extreme of both ranges; the support for the proposition that the rates are similar requires two rather poorly supported propositions to hold simulta,neously."

This flaw is often missed; see Wolfe R, Hanley J *If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2.* Canadian Medical Association Journal. 2002 Jan 8;166(1):65-6. [under **r e p r i n t s** in JH's main page].

• "The way to approach such problems [estimation of θ] is to reparametrize the model..."

---

[1]Morris JN, Marr JW, Clayton DG. Diet and heart: a postscript. Br Med J. 1977 Nov 19;2(6098):1307-14. During 1956-66, 337 healthy middle-aged men in London and southeast England participated in a seven-day individual weighed dietary survey. By the end of 1976, 45 of them had developed clinical coronary heart disease (CHD) which showed two main relationships with diet. Men with a high energy intake had a lower rate of disease than the rest, and, independently of this, so did men with a high intake of dietary fibre from cereals. Energy intake reflects physical activity, but the advantage of a diet high in cereal fibre cannot be explained; there was no evidence that the disease was associated with consumption of refined carbohydrates. Fewer cases of CHD developed among men with a relatively high ratio of polyunsaturated to saturated fatty acids in their diet, but the difference was not statistically significant.

Morris, who died in 2009, was an influential epidemiologist: see for example,
`http://www.ft.com/cms/s/2/e6ff90ea-9da2-11de-9f4a-00144feabdc0.html`
`http://en.wikipedia.org/wiki/Jerry_Morris_(physician)`
`http://ije.oxfordjournals.org/cgi/content/full/36/6/1184`

One could also first draw the iso-log-Likelihood (LL) contours of

$$LL(\lambda_0, \lambda_1) = D_0 \log(\lambda_0) - \lambda_0 Y_0 + D_1 \log(\lambda_1) - \lambda_1 Y_1,$$

over a grid of (λ_0, λ_1) values, then draw the lines $\lambda_1 = \theta \times \lambda_0$ as rays across this grid for various values of θ, say $\theta = \{2^{-2}, 2^{-1}, \ldots, 2^3\}$, say, and finally to find the maximum L over each line, and finally to plot these maxima as a function of θ.

13.2 Profile Likelihood

• "In most situations [the profile log likelihood] behaves in exactly the same way as a log likelihood".

Please check out the situations where it does not.

• "In the case of the the rate ratio, this process is simplified since the derivation of the profile log likelihood can be carried out algebraically, leading to a mathematical equation for the curve. The value of λ_0 which maximizes the log likelihood for any given value of θ may be shown to be..."

The full log-likelihood for (θ, λ_0) is

$$LL(\theta, \lambda_0) = D \log(\lambda_0) + D_1 \log(\theta) - \lambda_0 Y_0 - \theta\lambda_0 Y_1.$$

So, with θ fixed, its derivative with respect to λ_0 is

$$dLL(\theta, \lambda_0)/d\lambda_0 = D/\lambda_0 - Y_0 - \theta Y_1,$$

with its root at

$$\lambda_0 = D/(Y_0 + \theta Y_1).$$

[Incidentally, this solution makes intuitive sense, since we can think of $Y_0 + \theta Y_1$ as the number of 'reference-category-equivalent' person years: Y_1 person years at a (relative) rate of θ should produce as many events as θY_1 persons years would at a (reference) rate of 1. So, together, the combined Y_0 and Y_1 person years, in the reference and index categories respectively, can be expected to produce in total as many total cases as $Y_0 + \theta Y_1$ persons years in the reference category would.]

Upon substituting this expression for λ_0 into $LL(\theta, \lambda_0)$, and discarding terms that do not involve θ, we get, as C&H do, the Binomial-looking likelihood

$$LL_{profile}(\theta) = D_1 \log(\theta) - D \log(Y_0 + \theta Y_1).$$

This can be rewritten[2] as

$$LL_{profile}(\theta) = D_1 \log\left(\theta \times \frac{Y_1}{Y_0}\frac{Y_0}{Y_1}\right) - D \log\left(\frac{Y_0 + \theta Y_1}{Y_0}Y_0\right),$$

and the part that involves θ can be written as

$$LL_{profile}(\theta) = D_1 \log\left(\theta\frac{Y_1}{Y_0}\right) - D \log\left(1 + \theta\frac{Y_1}{Y_0}\right),$$

or as the binomial likelihood for the odds parameter Ω,

$$LL_{profile}(\Omega) = D_1 \log(\Omega) - D \log(1 + \Omega),$$

with

$$\Omega = \theta\frac{Y_1}{Y_0}.$$

- "From the Bernoulli likelihood, the most likely value of Ω is D_1/D_0."

This is their way of saying $\hat{\Omega}_{ML} = D_1/D_0$.

- "The standard deviation of $\log(\Omega)$ is ... ".

Again, they mean the standard deviation of $log\hat{\Omega}_{ML}$. We know from earlier in 601 – and the use of the two applications of the Delta method (first for $\frac{\hat{\pi}}{1-\hat{\pi}}$ and then for the log – that the variance of a logit, i.e. of a log-odds, is

$$Var\left[\log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right)\right] = Var[\log\hat{\Omega}] = Var[\hat{\pi}] \times \left(\frac{d\log\Omega}{d\Omega}\right)^2 \times \left(\frac{d\Omega}{d\pi}\right)^2$$

- "It follows that the most likely value of θ is ... "

$$\hat{\theta}\frac{Y_1}{Y_0} = \frac{D_1}{D_0} \quad \to \quad \hat{\theta} = \frac{D_1}{D_0} \div \frac{Y_1}{Y_0} = \frac{D_1}{Y_1} \div \frac{D_1}{Y_0} = \frac{\widehat{\lambda_1}}{\widehat{\lambda_0}}$$

"... which is the ratio of the most likely values of the two rates."

Supplementary Exercise 13.1 Score, score variance, and score test

Derive the expressions for the score and score variance, given at the bottom of C&H page 127. If you need to, consult C&H chapter 11.4, pp 102-104.

---

[2]JH thinks that there may be a typo (a minus instead of a plus) in C&Hs' addition.

13.3 Conditional Likelihood

- This follows directly from deciding to condition on the sum D of 2 independent Poisson random variables D_1 and D_0:

$$D_1 \sim Poisson(\mu_1); \; D_0 \sim Poisson(\mu_0) \to (D_1 \mid D) \sim Binomial\left(D, \pi = \frac{\mu_1}{\mu_1 + \mu_0}\right)$$

In our case, $\mu_1 = \lambda_1 Y_1$ and $\mu_0 = \lambda_0 Y_0$, so

$$(D_1 \mid D) \sim Binomial\left(D, \; \pi = \frac{\lambda_1 Y_1}{\lambda_0 Y_0 + \lambda_1 Y_1} = \frac{\theta Y_1}{Y_0 + \theta Y_1}\right).$$

Clearly, then,

$$\frac{D_1}{D_0} = \frac{\hat{\pi}}{1 - \hat{\pi}} = \frac{\hat{\theta}Y_1}{Y_0} \to \hat{\theta} = \frac{D_1}{D_0} \div \frac{Y_1}{Y_0} = \frac{D_1}{Y_1} \div \frac{D_0}{Y_0} = \frac{\widehat{\lambda_1}}{\widehat{\lambda_0}}.$$

So, in this instance, the conditional likelihood is exactly the same as the profile likelihood. And the ML estimators and the variances (approx. and exact) derived from curvature of the log-likelihood, also coincide.

- **Extra**: TOWARDS FITTING θ VIA A GENERALIZED LINEAR MODEL.

The above formulation of the odds $\Omega = \theta \times \frac{Y_1}{Y_0}$ in both the profile and conditional likelihoods lends itself to a regression approach to the fitting of θ. We have that

$$D_1|D \sim Binomial\left(D, \; \pi = \frac{\theta Y_1}{Y_0 + \theta Y_1}\right) \sim Binomial\left(D, \; \Omega = \theta\frac{Y_1}{Y_0}\right).$$

We can write this as

$$\frac{E[D_1]}{D - E[D_1]} = \theta\left(\frac{Y_1}{Y_0}\right),$$

or as

$$\log\left(\frac{E[D_1]}{D - E[D_1]}\right) = \log(\theta) + \log\left(\frac{Y_1}{Y_0}\right) = \beta \times X + 1 \times K.$$

where $\beta = \log(\theta)$ is a regression parameter to be estimated, $X = 1$, and $K = \log(Y_1/Y_0)$ is a known constant. In the language of generalized models, having $\log(\mu_1/[D - \mu_1])$ on the left hand side means that we are modeling the *logit* of $\mu_{Y|X}$, i.e., using the 'logit *link*', and we already have established that $D_1|D \sim Binomial(D, \; \pi)$, i.e., the '*distribution*' or the '*error* structure' or '*family*' is *Binomial*. Lastly, even though we represented the 'regression

equation' as having two terms, the $\beta \times X$ and the $1 \times K$, the *regression coefficient associated with K is known to be unity (1)*. In the GLM language, K is an '**offset**', i.e., it is a variate whose coefficient is known to be 1: we force this coefficient value to be 1, and we don't allow the software to estimate it.[3] Technically, it is part of the regression equation, but it is put in a separate part in the syntax of the sodtware. For example, in R, one would fit the above logistic model as follows:

```
D1=c(28); D=c(45); X=c(1); K=c(log(1857.5/2768.9));

fit=glm(cbind(D1,D-D1)~-1+X,family=binomial,offset=K)

        # format for grouped (Binomial) data
        # for Bernoulli (ie y=0/1) data, use y ~ -1+X etc

summary(fit)    [ Number of Fisher Scoring iterations: 3 ]

Deviance Residuals: [1]  0

Coefficients:
   Estimate Std. Error z value Pr(>|z|)
X   0.8982     0.3075   2.921  0.00349 **

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8.889  on 1  degrees of freedom
Residual deviance: 0.000  on 0  degrees of freedom
AIC: 6.2087

beta.hat=log.theta.hat=fit$coefficients
theta.hat=exp(log.theta.hat)
Var.beta.hat = summary(fit)$cov.unscaled[1,1]

c(beta.hat,theta.hat,Var.beta.hat)

0.898 2.455 0.094538

(28/1857.5)/(17/2768.9) = 2.455  #C&Hs' 2.48 slightly in error
```

```
1/28 + 1/17  = 0.09453782; sqrt(1/28 + 1/17) = 0.30747
```

The `glm` routine produces the same statistics as the hand fitting used in exercise 13.1.

We would have obtained the same output had we used the statement:

`glm(cbind(D1, D - D1) ~ 1, family = binomial, offset = K)`

---

[3]In this very simple example, with just one observation, K may look like an ordinary intercept, but remember that it is one whose value we *know*. Moreover, if there are several different observations, as there will be in Chapter 15, then K will vary from observation to observation, each time with a coefficient of 1.

Supplementary Exercise 13.1 The 1954 Field Trial of the Salk Poliomyelitis Vaccine[4]

Summary of Study Cases by Diagnostic Class and Vaccination Status (Rates per 100,000): **Placebo control areas: All Reported Cases**[*]

Gp.	SP	A_c	A_r	T_c	T_r	PP_c	PP_r	NP_c	NP_r	FP_c	FP_r
V	200,745	82	41	57	28	**33**	*16*	24	12	-	-
Pl	201,229	162	81	142	71	**115**	*57*	27	13	4	2
NI	338,778	182	54	157	46	**121**	*36*	36	11	-	-
IV	8,484	2	24	2	24	1	12	1	12	-	-
All	749,236	428	57	358	48	270	36	88	12	4	1

[*] *V, Pl, NI, IV: Vaccinated, Placebo, Not Inoculated, and Incomplete Vaccinations groups.*

SP: Study population (number of children);
A_c and A_r: *All* reported *cases* and *rate*;
T_c and T_r: *Total* poliomyelitis *cases* and *rate*;
PP_c and PP_c: *Paralytic Poliomyelitis cases* and *rate*;
NP_c and NP_r: *Non-Paralytic poliomyelitis cases* and *rate*;
FP_c and FP_r *Fatal poliomyelitis cases* and *rate*.
Some 70 reported cases were deemed to be "Not Polio" (25 in V, 20 in Pl, and 25 in the NI, are shown in Meier's table, but omitted here because of space constraints. Meier's Source: Adapted from Francis (1955), Tables 2 and 3.

The data are based on followup from the children (in grades 1 2 and 3), randomized and vaccinated in the months before the 1954 summer vacation, and followed to the end of December 1954; i.e. over one 'polio season'.

i. Repeat C&H's exercises 13.1 and 13.3 using the data of paralytic polio (PP) instead of those in C&H Table 13.1, i.e., compute point and interval estimates of the *difference* in, and *ratio* of, the rates of paralytic polio with the Salk vaccine and Placebo [first 2 rows].

Francis (1955) also used a conditional approach when computing their confidence intervals. See (on the top right corner of the BIO601 website, under Applications) the chapter on Statistical Methods. This chapter says...

With M =number of vaccinated persons; N = number of persons in corresponding control group; m = number of cases among vaccinated persons [treated as the realization of a Poisson random variable]; n =number of cases among persons in

corresponding control group [also treated as the realization of a Poisson random variable]; T =total number of persons ($T = M + N$); t = total number of cases ($t = m + n$), and $\lambda = RateRatio$, ...

the probability of m vaccinated cases, given a total of t cases, may be expressed as:

$$ {}^tC_m \times \left\{ \frac{M\lambda}{M\lambda + n} \right\}^m \times \left\{ \frac{N}{M\lambda + n} \right\}^{t-m} $$

This is the binomial distribution with probability parameter

$$ \frac{M\lambda}{M\lambda + N}. $$

ii. In this example, how similar would the CI have been if they had used an *unconditional* approach to the RateRatio? Is this similarity because of the large numbers of events, or because of something else?

iii. As is clear from the headline,[5] journalists, and the public, are more interested in the percent efficacy, $100 \times (1 - RateRatio)$, than in the RateRatio. Therefore, convert the CI for the RateRatio into a CI for the percent *efficacy*.

---

[4]Paul Meier. Chapter 2 The Biggest Public Health Experiment Ever: in Tanur JM et al. (Editors) *Statistics: A Guide to the Unknown.* Holden-Day San Francisco 1972.

[5]There are 3 virus strains; results above are not strain-specific

Supplementary Exercise 13.2 Women are Safer Pilots

> LONDON- Initial results of a study by Britain's Civil Aviation Authority shows that women behind the controls of a plane might be safer than men. The study shows that male pilots in general aviation are more likely to have accidents than female pilots. Only 6 per cent of Britain's general aviation pilots are women. According to the aviation magazine Flight International, there have been 138 fatal accidents in general aviation in the last 10 years, and only two involved women - less than 1.5 per cent of the total.

> Woman News, page F1 The Montreal Gazette, August 21st, 1995

In this example, since we do not know the absolute sizes of the 2 person-time denominators, and thus cannot compute absolute rates, or the difference in rates, we are limited to an analysis of the Rate *Ratio*.

The large-sample unconditional methods for obtaining a CI for a rate ratio are both tractable and accurate when there are enough events in each of the compared categories. But in this example, and in supp. exercise 13.3, the small number of events in one of the categories (2 in this example, 0 in the next example) renders large-sample methods inaccurate or even impossible. In such situations, the conditional approach, in which one bases the inference on the distribution of the number of events in one category, conditional on the sum of the numbers of events in the two categories, is one way around this problem (we use a similar conditioning strategy when dealing with Fisher's exact test). As C&H tell us at the top of p. 129, the profile likelihood – which has a binomial form – is the same as the conditional likelihood, and thus the same inference, in this type of example. But can we approximate this 'binomial likelihood' by a 'Normal approximation to binomial likelihood' (as C&H do in their example in Chapter 13[6] and still have accurate answers? In order to compare inferences based on an exact-conditional versus approximate-conditional approach to this particular example, we will compute interval-estimates for the ratio of the rates of accidents in women relative to men pilots using different approaches.

For simplicity, in what follows, assume that on average, women pilots fly just as many hours as the men pilots, and that all other relevant factors are equal, although they probably are not! – the true denominator ratio is probably more extreme than 6:94.

    i. Compute a 90% frequentist interval for the rate ratio θ by treating the

observed 2:136 split of the 138 cases as the realization of a binomial random variable, with $n = 138, \pi = \frac{6\theta}{6\theta+94}$. Do so by converting the 90%CI for π into a 90% CI for θ. Obtain the 90%CI for π in two ways:- (a) using an exact (Clopper-Pearson) CI for π (b) using a normal-approximation to the distribution of $\hat{\pi}$

  ii. Compute a 90% interval for the rate ratio θ using the Likelihood Ratio criterion of 0.258 i.e the Log Likelihood Ratio criterion of -1.353 [see C&H's remarks on page 22, and on pages 89-91].

Do so in 3 ways, using

    (a) the (binomial) likelihood $L(\theta)$,

    (b) the (binomial) likelihood $L(\log[\theta])$,

    (c) a normal-approximation to the (binomial) likelihood $L(\log[\theta])$

 iii. Compute a 90% credible interval for the rate ratio θ by treating the observed 2:136 split of the 138 cases as the realization of a binomial random variable, with $n = 138, \pi = \frac{6\theta}{6\theta+94}$. Again, obtain it in two ways:- using the binomial likelihood itself, along with a suitable (say a Beta) prior; using, as Spiegelhalter et al do, a normal-approximation to the likelihood, and a normal (i.e., Gaussian) prior.

 iv. Compute a 90% interval for the rate ratio θ using generalized linear model software, a logit link, and a single binomial observation (see example above). Which of the methods is the estimate from GLM equivalent to?

  v. (In point/bullet form] what are the take-home data-analysis messages from this exercise?

---

[6]The Spiegelhalter et al. text routinely uses a 'normal likelihood' (and 'normal prior').

Supplementary Exercise 13.3: A Controlled Trial of a Human Papillomavirus Type 16 Vaccine

Background: Approximately 20 percent of adults become infected with human papillomavirus type 16 (HPV-16).Although most infections are benign, some progress to anogenital cancer. A vaccine that reduces the incidence of HPV-16 infection may provide important public health benefits.

Methods: In this double-blind study, we randomly assigned 2392 young women (defined as females 16 to 23 years of age) to receive three doses of placebo or HPV-16 virus-like particle vaccine (40 ?g per dose), at day 0, month 2, and month 6. Genital samples to test for HPV-16 DNA were obtained at enrollment, one month after the third vaccination, and every six months thereafter. Women were referred for colposcopy according to a protocol. Biopsy tissue was evaluated for cervical intraepithelial neoplasia and analyzed for HPV-16 DNA with use of the polymerase chain reaction. The primary end point was persistent 16 infection, defined as the detection of HPV-16 DNA in samples obtained at two or more visits. The primary analysis was limited to women who were negative for HPV-16 DNA and HPV-16 antibodies at enrollment and HPV-16 DNA at month 7.

Results: The women were followed for a median of 17.4 months after completing the vaccination regimen. The incidence of persistent HPV-16 infection was 3.8 per 100 woman-years at risk in the placebo group and 0 per 100 woman-years at risk in the vaccine group (100 percent efficacy; 95 percent confidence interval, 90 to 100; P¡0.001).All nine cases of HPV-16related cervical intraepithelial neoplasia occurred among the placebo recipients.

Conclusions: Administration of this HPV-16 vaccine the incidence of both HPV-16 infection and HPV- 16related cervical intraepithelial neoplasia. Immunizing HPV-16negative women may eventually reduce the incidence of cervical cancer.

(N Engl J Med 2002;347:1645-51.). See full article on Resources-Applications webpage.

i. Why this design rather than a "fixed number of woman-years-of-follow-up" design?

ii. Let I denote incidence, v denote the vaccinated and u the unvaccinated, [Clayton and Hills use the general letters λ_v and λ_u]. Let IR denote the incidence ratio I_v/I_u, [Clayton and Hills use the general letter θ]. 'Vaccine efficacy' (E) is defined as a percentage

$$E = 100 \times (I_u - I_v)/I_u = 100 \times (1 - I_v/I_u) = 100 \times (1 - IR) = 100 \times (1 - \theta)$$

Consider a very large R.C.T., so that random variation is not an issue. A fraction F_v received the vaccine, and the average follow-up time was PT_V units; the remaining fraction $F_u = 1 - F_v$ received the placebo, and the average follow-up time was PT_U units. Denote the number of cases (of persistent infection) in the un-vaccinated sub-population by C_u and the corresponding number of cases in the vaccinated sub-population by C_v. Let $C_{total} = C_u + C_v$.

[Clayton and Hill use the letter D, presumably to stand for numbers of deaths; we use the more general letter C for 'cases of' , i.e., 'transitions' from the initial state (HPV-) to the state the vaccine is intended to prevent (persistent HPV+).] Let $\Pi = C_v/C_{total}$ denote the (theoretical) proportion of proportion of all cases that had been vaccinated.

• Assuming the rate ratio remains constant over time, express the parameter Π as a function of (a) the design parameters F_v and PT_v/PT_u, and the parameter IR (or θ), (b) the design parameters F_v and PT_v/PT_u, and the parameter E.

• Show the mathematical link between these parameters and the ones C&H use at the top of page 127.

iii. In the actual study cited above, the primary per-protocol efficacy analysis was based on observing persistent HPV-16 infection in 0 of 768 vaccinated women followed for 1084.0 woman years (w-y) and 41 in 765 unvaccinated women followed for 1076.9 women years (rate 3.8 per 100w-y).

This problem is similar in structure to that analyzed by C&H in their exercise 13.1 & 13.3. Examine the methods they used (their solutions are on p. 131-132.) Is it possible to use C&H's method with the HPV data?

If so, carry them out. If not, describe what other approach(es) is(are) possible, and carry it(them) out.

iv. (In point/bullet form) what are the take-home data-analysis messages from this exercise?

Supplementary Exercise 13.4: Extended Work Duration and the Risk of Self-reported Percutaneous Injuries in Interns

Refer to rows 2 and 3 of Table 3. in this article, by Ayas et al. in JAMA on Sept 6 of 2006. [Resources - Intensity]

i. Manually calculate ORs and 95% CIs, and repeat by computer software.

ii. Explain why your answers do not match those reported (hint: see the paragraph beginning "To assess the relationships..." in the last column of page 1057 of the article.

iii. exactly what (and how many) numbers would you need to carry out their analysis for row 3 (injuries in ICU). Answer in the form of a 1-paragraph request to the authors asking for these specific numbers (but do not e-mail the authors! JH has in fact obtained these numbers from Dr Ayas, and they will form the basis for some of a future homework).

iv. Is OR the correct term for the ratio being estimated here?