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 $(21.32)^2/30.37 = 14.97$, $(p < 0.001)$.

17 Likelihoods for the ~~Odds Ratio~~ Rate Ratio, i.e., Incidence Density Ratio, estimated with *estimated* *denominators*

PREAMBLE [JH]

The parameters being compared

How often is ***the*** odds ratio a parameter of direct scientific interest?

An odds ratio implies a comparison of two odds, i.e., of a mathematical transform of two proportions or probabilities.

Why and when to use, and to compare odds, probabilities, and probabilities per unit time[1]?

***

There are two situations in epidemiology where proportions arise naturally.

One is in the context of prevalence, where the *target* or *estimand* is say the per-cent or per-mille or per-million (prevalence) rate of being in a given life / health / illness / defect ***state*** at some specified time point. Examples are birth defects that are evident at birth, HPV seropositivity in male students beginning university, Facebook relationship status (JH isn't on Facebook, so he doesn't know if the there are more than two possibilities!) at the end of the first year in grad school, (un)employment, marital, children- and PhD status at age 30, professorial level (full vs associate/assistant) at age 45, undetected high blood pressure at age 50, retirement status, or still have all one's own teeth or hips or knees, at age 70, etc. In addition to these states, there are also ***traits***, such having a certain blood group or being a carrier of a good or bad gene.

Another, more biostatistical or technological, prevalences might be the proportion of R users who are using the various R versions (as of 2014-10-31, the latest is the Pumpkin Helmet version, R-3.1.2), or using the various Windows or MacOS or IOS or Ubuntu versions.

One reason not to focus on the prevalence itself, but on the prevalence odds, is that the log-odds is the natural (canonical) parameter in Bernoulli and Binomial regression, and that the parameters in an 'log-odds regression' (i.e., in a logistic regression) are – apart from their sign – the same whether we focus on $\omega = E[Y = 1]/E[Y = 0]$, or on $\omega = E[Y = 0]/E[Y = 1]$.

[1] C&H's term for a rate, what Rothman calls incidence, and what Miettinen calls incidence density

***

In the other situation, the *target* or *estimand* is the cumulative incidence, i.e. the prevalence rate, at the end of a given fixed period of time $[0, T]$, of *having made the transition* from an initial state $(Y = 0)$ at time $t = 0$ to the state of interest $(Y = 1)$ by (i.e. *at some time during* the interval from $t = 0$ to $t =] T$. Note that this is also a proportion, and as such it does not address how early or late within the fixed time interval the proportion was achieved, i.e. it is not concerned with the *speed* at which transitions occurred. As an example, if all we know is that at the end of January 2015, some 68% of iOS users are using iOS 8, we cannot infer how quickly or slowly the cumulative incidence has built to this level, since it was released at t_0 = September 17, 2014.

Examples of transitions within such a *fixed time interval* include deaths[2] within 30 days after birth, or after admission to a hospital, or an operation, or recurrence within 5 years after diagnosis of cancer, or CHD developing over the next 10 years in persons who are on say 50 years of age when they visit a doctor for their annual checkup.

In epidemiology, which seems to focus on transitions from good to bad states, the probability is usually called the **risk** over the time-span in question.

Note that a risk should always have a time-span attached to it. Otherwise, it has little meaning. In addition, the specification should include what, if they are substantial, is assumed about competing risks.

The absence of a specified time-span becomes even more problematic if we **compare** risks. In such comparisons, a ***risk ratio*** (also called a ***relative risk***) or a risk difference has even less meaning. For example, even though the mortality rates (rate of transitions from 'status: alive' to 'status: dead', i.e., incidence densities, or hazard rates, or forces of mortality, with $time^{-1}$ as their units), for men and women are approximately 1.4:1 across most of their lives, the ratio of their lifetime risks [age 0 to age 110, say] is 1.[3] Sadly, the term *relative risk*, with no time-span qualifier, is still widely used.

Non-epidemiological examples include the probability of (proportion) graduating within 5 years of entering a PhD program, or a first marriage by age 30, or becoming a grandparent, or retired, by age 65.

[2] death is an event, a transition from the 'alive' to the 'dead' state, i.e., a change of state; alive and dead are states, just as 'pre-op' and 'post-op' and 'post-MI' are.

[3] Economist John Maynard Keynes: 'But this long run is a misleading guide to current affairs. **In the long run we are all dead**. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again.

In these transition contexts, the concept of *sweeping across time* is central. Without the *passage of time*, there can be no transitions.

***

There is a **link between the prevalence proportion and the speed at which the transitions occur**. It involves the hazard function or the incidence density function, but the simplest version involves some steady state assumptions. Think of how the number of current students (prevalence) is related to (a) intake rates [incidence of new students] and (b) exit rates [or their reciprocals, the duration in the program].

The **link between the cumulative incidence proportion and the speed at which the transitions occur** [see equation below] involves, naturally enough, the hazard function or the incidence density function, a fundamental relationship we have already met.

***

Tighter definitions of these concepts are presented in the next 3 pages appended to these Notes. They are taken from an author who has thought more about these concepts than anyone else JH knows.

***

JH thinks that the **incidence density or hazard** as a more **fundamental parameter** (or function, if it is indexed by time or age) than the risk parameter is. For, the risk needs a time-referent, and there also needs to be clarity as to how competing risks are considered.

Moreover, there is a direct mathematical relation between the incidence density function, or hazard function $h(\cdot)$, over the time t' to t'' and the risk over this same period. This holds even in the absence of actual data where individuals have been followed up from time t' to t''. In other words, it is a *parametric* relation involving the incidence density function as a parametric function, of unknown form.

$$Risk_{t'\rightarrow t''} = 1 - \exp\left[-\int_{t'}^{t''} h(u)du\right] \qquad (1)$$

COMPARISONS of Prevalences, Risks, and Incidence Densities or Hazards

First, we need to start with the abstract, i.e. in terms of **parameters**, i.e, theoretically, before moving to data, and to the estimators of these comparative parameters.

Prevalences

In data-analyses that address comparisons of prevalences, or a prevalence function, it might be easier to use a ratio of odds rather than a ratio or difference of proportions to compare prevalences or fixed-horizon-risks in the index and reference categories of the contrast of interest. But even then, other comparative parameter such as or difference of proportions might be more meaningful.

Short-term Risks

A single incidence density or hazard can refer to a very short period of time (Δt), and if it does, the ratio of two such densities or hazards will be very close to the the ratio of their two associated risks over this short period. This is because, over a short period, so that the integral is small, and with $\overline{u}$ denoting the midpoint between 0 and Δt,

$$Risk_{1,0\rightarrow\Delta t} = 1 - \exp\left[-\int_0^{\Delta t} h_1(u)du\right] \approx 1 - \{1 - h_1(\overline{u}) \times \Delta t\} = h_1(\overline{u}) \times \Delta t$$

for the (0 *to* Δt) risk for the index category of exposure, with a corresponding one $Risk_{0,0\rightarrow\Delta t} \approx h_0(\overline{u}) \times \Delta t$ for the risk in the reference category of exposure.

> In this short-term, the risk ratio is just about equal to the ratio of the hazards,
>
> $$h_1(\overline{u})/h_0(\overline{u}).$$

Longer-term Risks

In the longer-term, suppose the index-category hazard function $h_1(t)$ is a constant (θ) times the $h_0(t)$ in the reference category, i.e. suppose we have 'proportional hazards', i.e.,

$$h_1(t) = \theta \times h_0(t) \ \forall \ t \text{ in the time-span of interest.}$$

Over most of adulthood, the age-specific mortality rates in males (1=M) vs. females (0=F), are more or less proportional, with a hazard ratio of approximately $\theta = 1.4$. If they are, what then is the ratio of the 10 year risks for males vs. females (a) aged 50 at the start? (b) aged 85 ? Letting G denote Gender,

$$Risk_{G,50\rightarrow 60} = 1 - \exp\left[-\int_{50}^{60} h_G(u)du\right]$$

Based on the force of mortality in the latest Canadian life tables, the 10-year integrals of the female hazard function from ages 50 to 60, and from 85 to 95, are approximately 0.03 and 1.18, respectively, so the 10-year risks are $1-\exp[-0.03] = 0.03$ and $1-\exp[-1.18] = 0.69$. For males, the corresponding integrals are 0.042 and 1.652, so that the 10-years risks are $1-\exp[-0.042] = 0.041$ and $1-\exp[-1.652] = 0.808$. Thus we have:

10 year age span	Risk: Male	Risk: Female	Relative Risk (M:F)
From 50 to 60	0.042	0.030	1.40
From 85 to 95	0.808	0.690	1.17

Again, in the short-term, with proportional hazards, the **risk ratio** is close to the ratio of the hazards.

But **in the longer-term, it is not**.

If indeed we have proportional hazards (i.e., the same $\theta = h_1(\cdot)/h_0(\cdot)$) across the full age or time range, the only easy-to-describe relationship that works well across that full range is the one involving the *surviving fractions*, i.e. the complements of the risks. In such situations, we have the simple relationship between the survival fraction in the index (1) and reference (0) categories

$$S_{1,t'\to t''} = [S_{0,t'\to t''}]^{\theta}.$$

In the examples above,

$$S_{M,50\to 60} = [S_{F,50\to 60}]^{1.4} = 0.97^{1.4} = 0.958,$$

and

$$S_{M,85\to 95} = [S_{F,85\to 95}]^{1.4} = 0.310^{1.4} = 0.192.$$

***

Comments on the various sections of Chapter 17 ...

> 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 ('retrospective' and 'prospective') 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.

Even though Figures 16.1 and 16.2 did involve some passage of time for the events to happen, ***both sections*** 1 (retrospective view, starting with cases, and 'comparing cases with controls' w.r.t. exposure) and 2 (prospective view, starting with exposure, and comparing exposed with unexposed' w.r.t. proportions of cases occurring in them) ***appear to ignore time***. It is as though there is a single 'instant' 2×2 table, with say exposure in the rows, and outcomes in the columns, that cross classifies people into the 4 cells by exposure and outcome, but without any time frame.

This would be fine if we were using a 2×2table to display the association between hair colour (natural, before aging, or use of hair dye) and eye colour, or outcomes where the passage of time is not central – e.g., acceptance of males versus females to medical school – or time is obscured, or – e.g., in sports – the lag between the shot and the goal is very short, or the states (normal, defect) occur at conception, and are determined by genetics or environmental factors, and there is no possibility of transition from the initial state.

In C&H's approach, as in Cornfield's in 1951, the prospective approach compares rows, the other one columns. Moreover, if one samples from row 1 with a sampling fraction f_{r1}, and from row 2 with sampling fraction f_{r2}, the log of the crossproduct ratio remains invariant to this. The same invariance holds if one one samples from column 1 with a sampling fraction f_{c1}, and from column 2 with sampling fraction f_{c2}, and indeed this sampling (usually with $f_{case.column} \gg f_{non.case.column}$) is what appears to be the defining feature of case control studies.

Just as Cornfield (1951)did, C&H show that, whichever way ones looks at it, whether one samples by row or by column, one obtains the same "odds ratio". They then proceed to use a **'2-binomials model'** to work out the sampling variability of the "odds ratio" estimator.

You will recognize the binomials by the variances of the logs of the two (identical) odds ratio estimators: In section 17.1 (retrospective, compare the exposure split of the cases ($D_1 : D_0$, one binomial) with exposure split of the

controls ($H_1 : H_0$, other binomial), it is

$$\underbrace{\frac{1}{D_1} + \frac{1}{D_0}} + \underbrace{\frac{1}{H_1} + \frac{1}{H_0}},$$

while in section 17.2 (prospective, compare the outcome split of the exposed ($D_1 : H_1$, one binomial) with the outcome split of the unexposed ($D_) : H_0$, other binomial), it is

$$\underbrace{\frac{1}{D_1} + \frac{1}{H_1}} + \underbrace{\frac{1}{D_0} + \frac{1}{H_0}}.$$

The 'rare disease assumption'

The loose justification for using the empirical 'odds ratio', as an estimator of a theoretical rate (incidence density) ratio (λ_1/λ_0) is given at the top of page 161 of C&H, but it is accompanied by two important assumptions that they freely admit are more likely to be violated than the 'rare disease assumption' that the derivation is based on.

This 'rare disease assumption' goes back at least as far as the classic 1951 Cornfield paper.[4] Unfortunately it still is the one given in many 'modern' texts, despite the much more general 'no need for rarity' derivation by Miettinen in 1976 (see Notes on Chapter 16).

Derivations that rely on the 'rare disease assumption' rest on algebraic arguments using *persons*, not *population time*. In C&H's derivation, the outcome *proportions* involved refer to *cumulative* incidence in some presumably fixed but unspecified span of time (they speak of 'failing over the period of the study' (p154, line 4). Throughout his 1951 paper, Cornfield did not even mention the passage of time: he speaks of prevalent cases, as though the occur instantly, from nowhere, and stay around (and thus can be studied) forever.

In section 16.5, C&H also start with proportions (risks), and show that the case-control design, carried out at the end of the study period, can estimate a ratio, not of the risks per se, but of their associated odds.

[4] A copy of the 1951 paper, accompanied by a commentary by Mitchell Gail, both taken from the 1997 book Breakthroughs in Statistics, can be found on the website.

THE MODERN OUTLOOK on so-called case-control studies

Later in section 16.5, C&H tell us that ***if we slice time very finely***, we do not need the rare disease assumption, or to have all persons followed for the same length of time, or to have no censoring. They provide us with a very simple way to do so.

> All of these assumptions can be guaranteed by the simple device of selecting a short enough study period. If insufficient cases would be obtained from such a study then the remedy is simple - carry out several consecutive short studies. The subjects remaining in the base at the end of one study immediately enter the next study. Each study then provides a separate estimate of the rate ratio, and provided this ratio remains constant over the whole study period, the information can be aggregated methods very similar to those discussed in Chapter 15.
>
> Taken to the limit, the total time available for the study may be divided into clicks which contain at most one case. Those clicks in which no case occurs are not informative so there is no purpose in drawing controls, but controls are drawn for all clicks in which a case occurs. Thus one or more controls are drawn from the study base immediately after the occurrence of each case. This design is termed incidence density sampling.
>
> A study carried out in this way involves matching of controls to cases with respect to time. Methods for stratified case-control studies will be discussed in Chapter 18, but in the special case where the ratio of exposed to unexposed persons in the study base does not vary appreciably over the study period, it is legitimate to ignore the matching by time during the analysis.

And so, we will from now on focus on this modern way of viewing case-control studies.

In this modern outlook, we think of the cases as arising in population-time, and we think of the population time involved as an infinite number of person-moments - think of a person-moment as a person at a particular moment. One way to represent population-time is to use the x-axis as time T and the y-axis as numbers of persons P, with P as a curve over T. Then population-time is the area of the surface under the P-T curve. If you shade this area using a very fine shading using a Laser printer, and zoom in on it, it might look like very fine dots (person moments) very close to together.

Obviously, in what follows, the time or age scale should not be so wide that the rates would vary too much from one end of it to the other. We typically work, as Miettinen did in the Table in the Notes for Ch 16, with age bands of at most 5 years of age, and calendar bands no more than a decade wide.

Say that , within a fairly narrow age and calendar band, a proportion π_1 of the population-time is "exposed" person moments, and the remaining proportion π_0 is "non-exposed" person-moments. Suppose further that the (theoretical) event rates in the exposed and unexposed amounts of population-time are

$$\lambda_1 = \frac{E[no.events]}{PT_1} \; ; \; \lambda_0 = \frac{E[no.events]}{PT_0}.$$

Estimands: $\theta = \lambda_1/\lambda_0$, and $\Delta\lambda = \lambda_1 - \lambda_0$

Estimation: It will always involve a **numerator** (case) series; As for **denominators** ...

i. If the absolute sizes of PT_E and PT_0 are known, we can estimate both $\theta = \lambda_1/\lambda_0$, and $\Delta\lambda = \lambda_1 - \lambda_0$ directly, as in chapters 14 and 15.

ii. If the absolute sizes of PT_1 and PT_0 are not known, but their relative sizes are, we can estimate θ directly, but not $\Delta\lambda$. Again, the methods of Chapter 14 and 15 apply, since only the offset, $\log(PT_1/PT_0)$, is needed.

iii. If $PT = PT_1 + PT_0$ is known, but its split is not, and if we obtain an unbiased estimate of this split, we can estimate both θ and $\Delta\lambda$.

iv. If nothing is known about the total PT or the exposure-specific PT's, but we are able to obtain an unbiased estimate of the $PT_1 : PT_0$ split, we can estimate just the ratio (θ) of the two λ's, but not their difference.

Data, and estimators

Numerator (Case) Series [overall size c; c_0, c_1 in 'exposure' categories 0, 1]

Denote by c the observed number of events; we classify them into c_1 events in "exposed" population-time and $c_0 = c - c_1$ in the "non-exposed" population-time. We will refer to this sample of c as the *case* series.[5]

If the absolute sizes of PT_1 and PT_0 are known, we can estimate both $\theta = \lambda_1/\lambda_0$, and $\Delta\lambda = \lambda_1 - \lambda_0$ directly, as in chapters 14 and 15.

$$\widehat{\lambda_i} = \frac{c_i}{PT_i}; \quad \widehat{\Delta\lambda} = \widehat{\lambda_1} - \widehat{\lambda_0}; \quad \hat{\theta} = \frac{c_1}{PT_1} \div \frac{c_0}{PT_0}.$$

If the absolute sizes of PT_1 and PT_0 are not known, but their relative sizes are, we can estimate θ directly, but not $\Delta\lambda$. Again, the methods of Chapter 14 and 15 apply, since only the offset, $\log(PT_1/PT_0)$, is needed.

$$\hat{\theta} \; = \; \frac{c_1}{PT_1} \div \frac{c_0}{PT_0} \; = \; \frac{c_1}{c_0} \div \frac{PT_1}{PT_0}.$$

If $PT = PT_1 + PT_0$ is known, but its split is not, and if we obtain an unbiased estimate of this split, we can estimate both θ and $\Delta\lambda$. To estimate the split, we rely on a

Denominator Series [random sample of the base from which the cases emerged, overall size d; with d_0, d_1 classified into 'exposure' categories 0, 1]. We will refer to this sample of d as the *denominator* series.

Then

$$\hat{\theta} = \frac{c_1}{\widehat{PT_1}} \div \frac{c_0}{\widehat{PT_0}} = \frac{c_1}{c_0} \div \frac{\widehat{PT_1}}{\widehat{PT_0}} = \frac{c_1}{c_0} \div \frac{d_1}{d_0}.$$

One can also use $\widehat{PT_i} = \frac{d_i}{d} \times PT$ to obtain $\widehat{\lambda_i}$ and $\widehat{\Delta\lambda}$.

[5] In this section, JH has borrowed from Miettinen the notation of 'c_1' and 'c_0' for the numbers of exposed and unexposed cases, and 'd_1' and 'd_0' for the numbers of exposed and unexposed denominators, instead of C&H's D_1 and D_0, and Y_1 and Y_0. Both of these sets of notations are more memorable and instructive than the A, B, C, and D in Mantel and Haenszel's 1959 paper, and the a, b, c, d also widely used elsewhere is statistics.

If nothing is known about the total PT or the exposure-specific PT's, but we are able to obtain – from a suitable denominator series – an unbiased estimate of the $PT_1 : PT_0$ split, we can estimate just the ratio (θ) of the two λ's, but not their difference.

$$\hat{\theta} = \frac{c_1}{\widehat{PT_1}} \div \frac{c_0}{\widehat{PT_0}} = \frac{c_1}{c_0} \div \frac{\widehat{PT_1}}{\widehat{PT_0}} = \frac{c_1}{c_0} \div \frac{d_1}{d_0}.$$

Statistical models for the estimators

• What is the *statistical model* for the $c_1 : c_0$ split?

We can think of c_1 as the realization of a Poisson r.v. with mean (expectation) $\mu_1 = (PT \times \pi_1) \times \lambda_1$. Likewise, think of c_0 as the realization of a Poisson r.v. with mean (expectation) $\mu_0 = (PT \times \pi_0) \times \lambda_0$.

Now, it is a statistical theorem (Casella and Berger, p194, exercise 4.15) that

$$c_1 \mid c \sim Binomial\left(c,\ \pi = \frac{\mu_1}{\mu_1 + \mu_0}\right);\ \frac{\pi}{1-\pi} = \frac{PT \times \pi_1 \times \lambda_1}{PT \times \pi_0 \times \lambda_0} = \frac{\pi_1}{\pi_0}\theta.$$

• What is the *statistical model* for the $d_1 : d_0$ split? Clearly, it is

$$d_1 \mid d \sim Binomial(d, \pi_1).$$

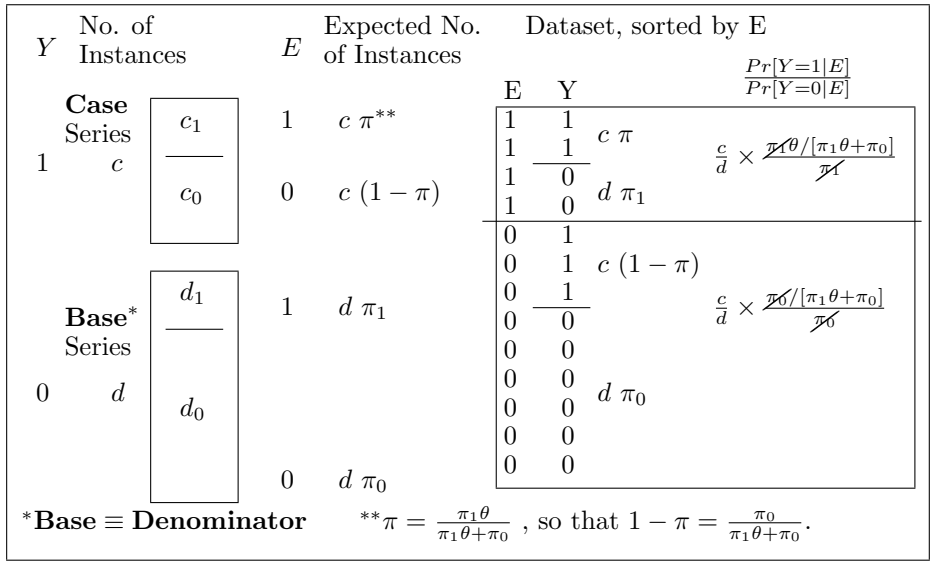
Coupling both series so as to estimate $\log\theta$

The $c_1 : c_0$ split is governed by **one binomial**, *involving* θ and π_1/π_0, while the $d_1 : d_0$ split is governed by **a separate binomial**, involving the same other parameter π_1/π_0, but *not* involving θ.

Now, consider the dataset of $c + d$ observations, with each of the c observations in the case series coded as $Y = 1$, and each of the d observations in the denominator series coded as $Y = 0$, and with each of the $c_1 + d_1$ observations in the combined series coded as $E = 1$, and each of the remaining $c_0 + d_0$ observations in the combined series coded as $E = 0$. The dataset,and how it came to be (i.e. starting at the top left with the case series, then adding a denominator series) is shown in the **diagram**.

To be parsimonious, and allow for a GLM approach, we can combine the two models into one 'master' regression equation.

In the diagram, to the right of the $E = 1$ portion of the dataset, the expression for the expected $Y = 1 : Y = 0$ split (i.e. the odds) is given. Taking the log of the odds, we get



$$\text{logit} = \log\left\{\frac{Pr[Y=1|E=1]}{Pr[Y=0|E=1]}\right\} = \log\left\{\frac{c}{d}\right\} + 1 \times \log[\theta] - \log[\pi_1\theta + \pi_0]$$

Similarly, taking the log of the odds in the $E = 0$ portion of the dataset,

$$\text{logit} = \log\left\{\frac{Pr[Y=1|E=0]}{Pr[Y=0|E=0]}\right\} = \log\left\{\frac{c}{d}\right\} + \cancel{0 \times \log[\theta]} - \log[\pi_1\theta + \pi_0]$$

The difference in logits is $\log(\theta)$. So, if we have a regressor variate that turns off $\log(\theta)$) when the variate is set to 0, and that turns it on when it is set to 1, we should be able to extract a $\widehat{\log(\theta)}$ from a logistic regression model fitted to the dataset.

So, one can estimate $\log\theta$ by an unconditional logistic regression of the Y's ($c + d$ observations in all, with $Y = 1$ if in case series; $= 0$ if in denominator series) on the corresponding set of $c + d$ indicators of exposure (E = 1 if exposed, 0 if not).

Even though it was assembled by starting with the case series, this dataset looks like it was assembled in the usual prospective manner, and it is very much in the modern spirit of comparing event rates in the exposed and unexposed population time.

We now proceed to the parameter fitting....

Fitting $\log\theta$ by GLM

```
# data entered in order of arrival, case series first
   y = c(     rep(1,5) ,           rep(0,8))
   E = c( rep(1,2),rep(0,3) , rep(1,2), rep(0,6))

ds = data.frame(y,E) ;retrospective = ds;
names(retrospective)=c("I.case.r","I.exposed.r")
prospective = ds[order(-E,-y),2:1]
names(  prospective)=c("I.exposed.p", "I.case.p")

cbind(ds,retrospective,prospective ) ;
   y E I.case.r I.exposed.r I.exposed.p I.case.p
1  1 1        1           1           1        1
2  1 1        1           1           1        1
6  1 0        1           0           1        0
7  1 0        1           0           1        0
3  1 0        1           0           0        1
4  0 1        0           1           0        1
5  0 1        0           1           0        1
8  0 0        0           0           0        0
9  0 0        0           0           0        0
10 0 0        0           0           0        0
11 0 0        0           0           0        0
12 0 0        0           0           0        0
13 0 0        0           0           0        0

### fitting via 'prospective' model

fit.pro = glm(I.case.p~I.exposed.p,family=binomial,data=prospective)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6931     0.7071  -0.980    0.327
I.exposed.p   0.6931     1.2247   0.566    0.571

exp(fit.pro$coefficients) : (Intercept) I.exposed.p
                                  0.5         2.0
round(summary(fit.pro)$cov.unscaled,2);  # Var[par. estimates]
                           (Intercept) I.exposed.p
               (Intercept)         0.5        -0.5
               I.exposed.p        -0.5         1.5
```

```
>    round(fit.pro$fitted.values,2)
   1    2    6    7    3    4    5    8    9   10   11   12   13
0.50 0.50 0.50 0.50 0.33 0.33 0.33 0.33 0.33 0.33 0.33 0.33 0.33
>
> aggregate(fit.pro$fitted.values,
+           by=list(E=prospective$I.exposed),sum)
  E x
1 0 3
2 1 2
> aggregate(1-fit.pro$fitted.values,
+           by=list(E=prospective$I.exposed),sum)
  E x
1 0 6
2 1 2

# If use log(b/d) =  log(size of case series/ size of base series)
#  = log(5/8) as a (common) offset, what does intercept denote?

OFFSET = rep( log(5/8), 13)

fit.pro.with.offset = glm(I.case.p ~ offset(OFFSET) + I.exposed.p,
          family=binomial,data=prospective)
summary(fit.pro.with.offset)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2231     0.7071  -0.316    0.752
I.exposed.p   0.6931     1.2247   0.566    0.571
```

Supplementary Exercise 17.1p

Using the diagram and logit equations (p.6) as a guide, *interpret* the two fitted coefficients from the GLM fit of the prospective model. Do not take *interpret* to mean stating whether they are 'statistically significant'; instead, tell us what *epidemiologic parameters* these coefficients are estimating (are estimates of), (a) 'as is', (b) when exponentiated and (c) further back transformed.

Display the sums of the fitted frequencies as a 2×2 table ($E \times y$) of fitted frequencies. Insert these into Woolf's formula and calculate $\text{Var}_{Woolf}[\widehat{\log\theta}]$. Compare your result with the GLM output. Comment.

In the model that includes the $\log 5/8$ as a common offset, what parameter (amalgam) is the -0.2231 an estimate of? Hint: Cf the logit equations. Is the intercept with the offset removed any more interpretable than the one that ignores the relative sizes of c and d?

Supplementary Exercise 17.1r

```
## retro

> fit.retro = glm(I.exposed.r~I.case.r,family=binomial,
   data=retrospective)
> summary(fit.retro)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0986     0.8165  -1.346    0.178
I.case.r      0.6931     1.2247   0.566    0.571

    Null deviance: 16.048  on 12  degrees of freedom
Residual deviance: 15.727  on 11  degrees of freedom

> exp(fit.retro$coefficients)

(Intercept)    I.case.r
       0.33         2.0
>
> round(summary(fit.retro)$cov.unscaled,2)
            (Intercept) I.case.r
(Intercept)        0.67    -0.67
I.case.r          -0.67     1.50
>
>    round(fit.retro$fitted.values,2)  # fitted E[y]'s

   1    2    3    4    5    6    7    8    9   10   11   12   13
0.40 0.40 0.40 0.40 0.40 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25
>
> aggregate(fit.retro$fitted.values,
+           by=list(y=retrospective$I.case.r),sum)
  y x
1 0 2
2 1 2
> aggregate(1-fit.retro$fitted.values,
+           by=list(y=retrospective$I.case.r),sum)
  y x
1 0 6
2 1 3
```

Focus on the left hand side (the retro version) of the diagram on page 6.

• Derive the log odds equation, involving $\frac{Pr[E=1|y=1]}{Pr[E=0|y=1]}$, for the c split;

do the same for the corresponding one, involving $\frac{Pr[E=1|y=0]}{Pr[E=0|y=0]}$, for the d split.

BTW: You should find the algebra needed to derive the logit equation for the d split involves just one parameter, and that it does not involve the parameter of interest, $\theta = \lambda_{E=1}/\lambda_{E=0}$.

• Using these 2 retrospective logit equations as a guide, *interpret* the two fitted coefficients from the GLM fit of the retrospective model. Again, do not take *interpret* to mean stating whether they are 'statistically significant'; instead, tell us what *epidemiologic parameters* these coefficients are estimating (are estimates of), (a) 'as is', (b) when exponentiated and (c) further back-transformed.

• Display the sums of the fitted frequencies as a 2×2 table ($y \times E$) of fitted frequencies. Insert these into Woolf's formula and calculate $\text{Var}_{Woolf}[\widehat{\log\theta}]$. Compare your result with the GLM output. Comment. Which variance version in C&H (page 167 or page 170) does this version correspond to?

• In which of the two approaches, retrospective, or prospective, is the intercept a more interpretable parameter estimate, and why ? In which of the two approaches, retrospective, or prospective, is the other coefficient a more interpretable parameter estimate, and why ?

• Which of the two approaches, retrospective, and prospective, do you prefer, and why ?

Supplementary Exercise 17.2

Refer again to the Danish study on the possible role of MMR vaccination on the aetiology of autism, and in particular, for this exercise, focus on just 1 square, the 1993 birth cohort, in the children time between their 3rd and 4th birthdays. This child time consisted of some Y_1=60,143 vaccinated child-years, and Y_0=6,857 unvaccinated child-years.

Using your counts of the numbers (c_1 & c_1) of cases in these vaccinated ($\cdot_1$) and unvaccinated ($\cdot_0$) child-years, and the frequencies (d_1 & d_1) in your simulated denominator series of size $d = 100$, set up an 'individual-record' datafile (of length $c + d$). You can use the toy example, in particular the right hand side of the in the diagram on page 6 and the R code on page 7, as a template. Then use logistic regression to obtain a point and interval estimate of $\log\theta$ and of θ itself.

Do you trust the Gaussian-based interval estimate? Why/why not?

Supplementary Exercise 17.3

Refer again to the article by Woolf (1955), and restrict your attention to the London portion of the contrast in his Table 1, ie. to the data in the first row of the Table.. Both directly, and by using logistic regression[6] calculate a point and an interval estimate for what Woolf calls x and what we call θ.

Verify that the logistic regression produces the 'Woolf' variance.

Supplementary Exercise 17.4[7]

Woolf's estimator of θ is, in his notation,

$$\hat{\theta} = \frac{h}{H} / \frac{k}{K} = \frac{hK}{kH} \quad \left[\text{ or } \frac{ad}{bc} \text{ in epidemiologists' notation} \right],$$

where h & k are from the case series, and H & K are from the denominator (base) series.

Treat $H : K$ as a binomial sample of size $H+K$ from a base where the fractions of the overall population time that are the index and reference categories of the contrast are π_1 and $\pi_0 = 1 - \pi_1$ respectively. In other words, assume that the base is effectively infinite in size, relative to $H + K$, so that even if, as one would, one samples without replacement, we can treat $H \sim$ Binomial$(H + K, \pi_1)$.

• Use the delta method[8] to obtain the (approx., large sample) variance for $\log[H/K]$. Use the observed values of H of K as plug-in estimates for the E$[H]$ and E$[K]$ in this theoretical variance.

Now, treat h and k as two independent Poisson random variables, representing the numbers of cases arising from these same index and reference categories of the base. Take advantage of the fact that we can treat $h \sim$ Binomial$(h + k, \pi)$,[9] and of the variance form you just obtained for the log of a ratio in the denominator series.

• By doing so, obtain the (approx., large sample) variance for $\log[h/k]$ Again, as above, use the observed values of h and k as plug-in estimates for the E$[h]$ and E$[k]$ in this theoretical variance.

• Now, use these 2 variances to obtain an expression for $\log[\hat{\theta}] = \log[\frac{hK}{kH}]$. You should get the same variance formula that Woolf used.

[6] Cf. the template from the toy example of logistic regression on pages 6 and 7.

[7] a favourite in the comprehensive exam

[8] or the Fisher Information from a binomial likelihood for the parameter $\log \frac{\pi}{1-\pi_1}$,

[9] π is now an amalgam, whose form need not concern you for now, of the parameters π_1 and θ.

Remarks re Supplementary Exercise 16.4

In our simplified and fictional example, there had been 7 airline crashes involving planes from 2 of the leading manufacturers A and B. Using the denominator totals ?? :??, we tallied in class, you could calculate a (crude) Rate Ratio estimate $\frac{5B/2A}{33B/45A}$ for the B:A contrast, along with an interval estimate.

Setting aside all the other shortcomings of our fictional study, you must wonder if the small numerators make the use of a Gaussian-based interval inappropriate. If all you wanted was a p-value to judge the evidence against the null, you could always use Fisher's exact test, which is based on the central (null) hypergeometric distribution, which conditions on all four table margins. But what if you wanted a point estimate and interval estimate for the Rate Ratio (of course it would be wide) that made use of an exact distribution, rather than a Gaussian approximation. You could look up the use of the non-central hypergeometric distribution in Chapter 4 of Breslow and Day Volume I, along with their nice small worked example on page xx. You are not asked to repeat the calculations with our A:B crash data, but please be aware of the conditional approach to point and interval estimation. It is not as big a computational deal as it was when B&D was written. And this conditional likelihood also comes into the fitting of the parameters of the Cox model.

Supplementary Exercise 17.5

This exercise is designed to reinforce the ideas in the following quote from Mantel's 1973 paper on synthetic' studies. He had too many instances of $Y = 0$ in relation to the numbers of instances of $Y = 1$, and not enough computing resources, and so he sampled from the $Y = 0$ instances.

Focus on the portion *in italics* (added by JH) in the following excerpt.

> If we chose π_1 as 1 and π_2 as 0.15, we would have all the cases and 3.5 negatives per case. By the reasoning that $n_1 n_2/(n_1 + n_2)$ measures the relative information[10] in a comparison of two averages based on sample sizes of n_1 and n_2 respectively, we might expect by analogy, which would of course not be exact in the present case, that this approach would result in only a moderate loss of information. *(The practicing statistician is generally aware of this kind of thing. There is little to be gained by letting the size of the control group, n_2, become arbitrarily large if the size of the experimental group, n_1, must remain fixed.)*
>
> But the reduction in computer time would permit much more effective analyses. Ostensibly we would be meeting the additional conditions assumed for validity of the retrospective study approach; that is the retained individuals would be a random sample of the cases and disease-free individuals arising in the prospective study.... p.481.

[10] Note that $n_1 n_2/(n_1 + n_2)$ is algebraicly equivalent to the reciprocal of $1/n_1 + 1/n_2$. The product of $(1/n_1 + 1/n_2)$ and the square of the within-group SD is the square of the SE of the difference of two averages.

i. Use the data from the London portion of the data in Woolf's Table 1, and the R code provided on the Website, to numerically illustrate how "little (is) to be gained" by having the denominator series become arbitrarily large: Calculate what the variance (and its reciprocal, the amount of '*information*') would have been if the denominator series were (i) the same (ii) twice (iv) four times (x) 10 times (c) 100 times and (m) 1000 times the size (as)of the numerator series. [11]

   Using the amount of information in the'1000 times as large' scenario as the *maximum* information I_{max}, calculate and plot the ratio of each of the other amounts of information, I_{lesser}/I_{max} as a function of the ratio of the size of the denominator series to the size of the numerator series. *The computations can be carried out using the R code provided.*

ii. Can you see from your plot why a 'control:case' ratio of 4 has come to be regarded as a compromise? Can you think of situations where this magic number doesn't always work? As one possible situation, *simulate* various size denominator series sampled from the 40,046 homes known to have been supplied by the Southwalk and Vauxhall Water Company and the 26,107 known to have been supplied by the Lambeth Company (in the 300 homes in the numerator ('case') series, the split was 286:14). In this example, why is there so little extra gained even in going from 1:1 to 2:1? *Hint: find the 'weakest link' in the Woolf variance for the log(idr) estimate.*

iii. Suppose the denominators of 40,046 and 26,107 homes (and thus known person-time denominators) are *known and without error*. Calculate the variance of the log(idr) and use it to superimpose a 95% CI for the IDR on the graph produced in Question 6.

   What would the Woolf variance have been if Snow were forced to *estimate* the ratio of the number of homes supplied by the Southwalk and Vauxhall Water Company to the number supplied by the Lambeth Company, using a 'denominator-series' of a total of just 100 homes, and an observed split of 65:35?

   Treat the ratio of the denominators (O and A, called H and K by Woolf) in Woolf's example as *random-error-containing* estimates of the true ratio of the two person-time denominators that gave rise to the numerators o and a (called h and k by Woolf).

   What is the formula for the variance of the log(idr)? What is the role of the $(1/O+1/A)$ term in this variance formula? the role of the $(1/o+1/a)$ term?

[11] Scale the observed frequencies in the denominator series accordingly – in practice, because of sampling variation, they would not scale exactly.

iv. Use the comparison of the variance formula in which the denominator ratio has to be estimated with the variance formula in which the denominator ratio is treated as known, as a motivation for how to explain the main data-analysis difference between so called 'case-control' and 'cohort' studies:

> "When estimating an incidence density ratio, the main data-analysis difference between doing so in a so-called 'case-"control" ' and in a'cohort' study is that in the _____ study the person-time denominators are _____ whereas in the _____ study the person-time denominators are _____.

Supplementary Exercise 17.6

Refer to table I of, Miettinen's 1976 paper, where he computes age-specific Incidence Density Ratios, and 30 year risks of bladder cancer for smokers and non-smokers.

This paper showed that the rare disease assumption is not needed in 'case-control' studies, i.e., studies that sample the base (the PT experience in which the cases arose) to estimate the denominator ratios, and obtain rate ratio estimates. If, as in Table 1, the overall PT in each stratum is known, the absolute rates (and their differences) can also be estimated.

Note also the calculation of 30-year risks, and the ratio of these 30-year risks.

Using the R code provided, or otherwise, and without mentioning the 'OR' phrase, ...

i. Calculate a summary Rate Ratio (point and interval) estimate, using Woolf's method

ii. Calculate a summary Rate Ratio point estimate, using the Mantel-Haenszel method. Find a formula for the variance of this (calculations themselves not needed)

iii. How many strata were there in the 1959 Mantel-Haenszel worked example (Cf. link: table is on p738/9)? Were the M-H data sparser/more abundant than the Miettinen data?

iv. Obtain a Rate Ratio interval estimate via logistic regression.

v. Calculate 30 year risks (in the absence of competing causes) of bladder cancer for smokers and non smokers.

Supplementary Exercise 17.7

Refer to the numerator series JH has compiled from the data you extracted for subsets of the Canadian-born NHL players.

The ***parameter*** of interest is the 'Rate of reaching the NHL' as a smooth function of birth month (01-12).

Using the R code provided, or otherwise, and without mentioning the 'OR' phrase, ...

i. Fit a suitable rate function, using the population-based denominators[12]. Compare the fitted rates for January- and December-born Canadians, and obtain a CI for the ratio.

   [Note: the R code reads in denominators from a .csv file that is also on the website.]

ii. Fit a suitable rate function, using the sample of current-day Senators and MPs (Cf. Website) as a 'denominator series.' Compare the SE's of the parameters fitted to these data with those obtained from the population-based ones, and if they are very different, explain why they differ.

SAMPLE SIZE/POWER CONSIDERATIONS FOR (UNSTRATIFIED) 'CASE-CONTROL' STUDIES

In what follows, the parameter of interest is the Rate Ratio ($\theta = \lambda_1/\lambda_0$) estimated by a cross-product ratio involving numerators and estimated denominators. These numerators and denominators can be thought of as arising from 2 binomial models. The binomial for the denominator series is governed by just the proportions π_1 and $\pi_0 = 1-\pi_1$ of exposed and unexposed and unexposed population time. The one for the numerator series is governed by a more complex parameter π that is a function of that same parameter π but also the Rate Ratio parameter θ.

$$\pi = \frac{\theta\pi_1}{\theta\pi_1 + (1-\pi_1)}$$

Testing $H_0: \theta = 1$ vs. $H_{alt}: \theta \neq 1$ is equivalent to testing that

$$H_0: \pi = \frac{1\times\pi_1}{1\times\pi_1 + (1-\pi_1)} = \pi_1 \quad vs. \quad H_{alt}: \pi \neq \pi_1.$$

[12] admittedly, they are too recent, but patterns have not changed dramatically over the decades, and it would take too long to get better ones, closer to the players' years of births.

c & d for power $1-\beta$ if $\theta = \theta_{alt}$; Prob[Type I error] $= \alpha$.

Work in $\log\theta$ scale, so that

$$SE[\widehat{\log\theta}] = (1/c_1 + 1/c_0 + 1/d_1 + 1/d_0)^{1/2}.$$

Need

$$Z_{\alpha/2}\, SE_0[\widehat{\log\theta}] + Z_\beta SE_{alt}[\widehat{\log\theta}] < \Delta.$$

where

$$\Delta = \log[\theta_{alt}]$$

Substitute expected c_1, c_0, d_1, d_0 values under null and alt. into SE's and solve for c and d.
References: Schlesselman, Breslow and Day, Volume II, ...

Key Points: $\widehat{\log\theta}$ most precise when c_1, c_0, d_1, d_0 of equal size; so,

i. increasing the case series:base series ratio (ie., control:case ratio) leads to diminishing marginal gains in precision.

   To see this... examine the function

$$\frac{1}{\#\ of\ cases} + \frac{1}{multiple\ of\ this}$$

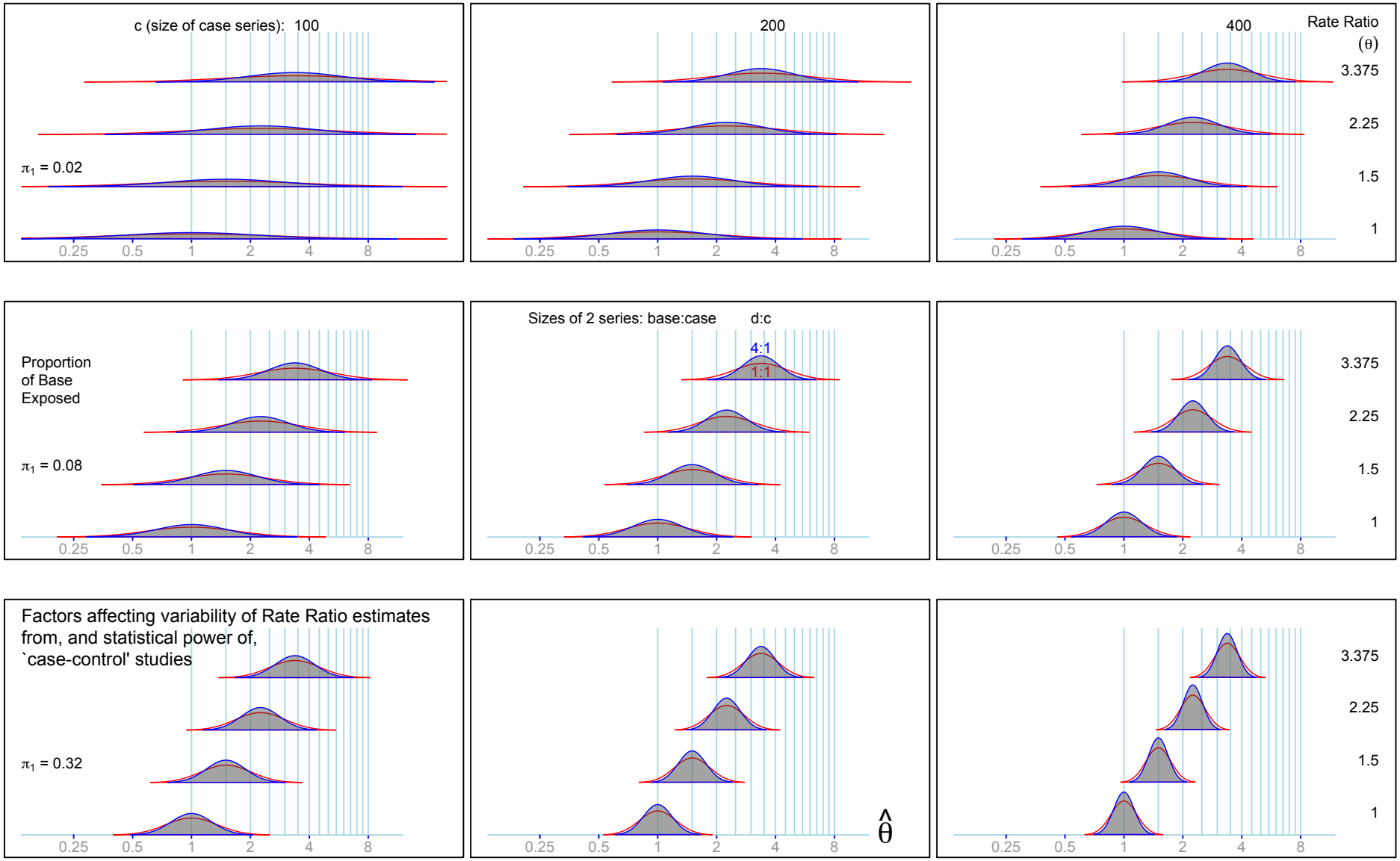
   for various values of "multiple"

ii. The more unequal the distribution of the etiologic / preventive factor, i.e., the more extreme the $\pi_1 : (1-\pi_1)$ split, the less precise the estimate

   Examine the functions

$$1/d_1 + 1/d_0 \quad \& \quad 1/c_1 + 1/c_0.$$

See, e.g., middle panel of **graph overleaf**, with log scale for $\hat\theta$: $c = 200$ cases, and exposure prevalence $\pi_1 = 8\%$. Say the Type I error rate set at $\alpha = 0.05$ (2-sided) so that upper critical value (that cuts off top 2.5% of null distribution) is close to $\hat\theta = 2$. Draw vertical line at this critical value, and examine how much of each non-null distribution falls to the right of it. This area to the right of the critical value is the **power** of the study, i.e., the probability of obtaining a significant $\hat\theta$, when in fact the indicated non-null value of θ is correct. The two curves at each θ value are for studies with $d:c$ = 4:1 and 1:1.



Power is larger if ...
(a) non-null $\theta \gg 1$ (Cf. 1.5 vs. 2.25 vs 3.375);
(b) exposure common (cf. 2% vs. 8% vs. 32%) but not 'too' common;
(c) larger c (Cf. 100 vs. 200 vs. 400), and $d : c$ ratio (1 vs. 4).