15
Comparison of rates within strata

15.1 The proportional hazards model

Direct standardization is a very simple way of correcting for confounding but it does have some limitations. This chapter deals with the alternative and more generally useful approach of stratification. We shall again illustrate our argument using the study of the relationship between energy intake and IHD first introduced in Chapter 13 and further analysed in Chapter 14. There, in Table 14.1, we showed the data stratified by 10-year age bands and demonstrated that the low energy intake group is, on average, rather older. This might explain some, or all, of the increase in IHD incidence rate. The method of direct standardization predicts the marginal rates for energy intake groups with the same standard age distribution. This chapter explores the alternative approach which compares age-specific rates within strata. Table 15.1 extends Table 14.1 by calculating rate ratios within each age band. This demonstrates the main problem with this approach to confounding; holding age constant and making comparisons within age strata leads to variable and unreliable estimates, because the age-specific rates are based on so few data.

This problem is resolved is by combining the age-specific comparisons from the separate strata, but any such procedure carries with it a further modelling assumption, because combining the age-specific comparisons can only be legitimate if we believe that they all estimate the same underlying quantity. If we are prepared to believe that the rate ratio between exposure

Table 15.1. Rate ratios within age strata

	Exposed (< 2750 kcal)			Unexposed (≥ 2750 kcal)			Rate
Age	D	Y	Rate	D	Y	Rate	ratio
40–49	2	311.9	6.41	4	607.9	6.58	0.97
50–59	12	878.1	13.67	5	1272.1	3.93	3.48
60–69	14	667.5	20.97	8	888.9	9.00	2.33
Total	28	1857.5	15.07	17	2768.9	6.14	2.45

groups is constant across age-bands, the evidence from the three bands can be brought together to provide a single estimate of the (constant) age-specific rate ratio. Of course the model on which the estimate is based, like all models, is open to question and in later chapters we shall discuss ways in which we can test whether it holds. For the present, we shall be content to believe that the model holds in our example, and that the fluctuation of age-specific rate ratios in Table 15.1 is no more than we would expect given the small numbers of cases in each age band.

Our notation follows naturally from earlier chapters. The age bands are indexed by the superscript t and exposure groups are indexed by subscripts, so that λ_0^t and λ_1^t are the rate parameters in age band t for the unexposed and exposed subjects respectively. We shall write the rate ratio parameter as θ, so that the model of constant rate ratio may be written

$$\frac{\lambda_1^t}{\lambda_0^t} = \theta.$$

This is called the *proportional hazards* model. The parameter θ is called the rate ratio for exposure *controlled for* age, sometimes abbreviated to the *effect* of exposure controlled for age. In this chapter we discuss how θ can be estimated.

15.2 The likelihood for θ

When the rate ratio is constant across age bands, we can replace the rate parameters λ_1^t by $\theta\lambda_0^t$. In our example, this reparametrization replaces the original six rate parameters, which we assume to be constrained to obey the proportional hazards model, with four parameters which are free to take any positive value. One parameter, namely the rate ratio θ, is our prime interest, and the remaining three are regarded as nuisance parameters.

Since each age band serves as an independent study, it is a simple matter to write down the log likelihood for a stratified comparison. Constructing the log likelihood using the prospective argument, each age band contributes a term which depends upon θ and the appropriate λ_0^t. The total likelihood is obtained by adding these terms over age bands. For comparing rates between exposed and unexposed subjects, the parameters λ_0^t are nuisance parameters. As in Chapter 13, replacing these by their most likely value for given θ leads to a profile log likelihood for θ. With the caveat expressed at the end of section 13.3, this log likelihood can also be justified as a conditional likelihood based on the split of cases within each stratum.

The log likelihood ratio curve for $\log(\theta)$ in our illustrative example is shown in Figure 15.1. Using a computer, it is a simple matter to find the most likely value, M, and to use the curvature of the log likelihood ratio to compute a Gaussian approximation. In this case $M = 0.8697$

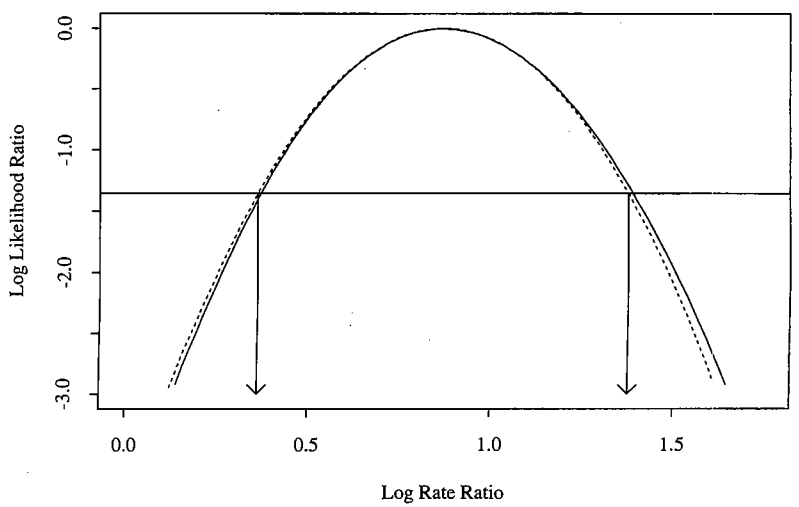


Fig. 15.1. Log likelihood ratio for the common rate ratio.

and $S = 0.3080$, and this approximation is shown as a broken line in the figure. The most likely value of the rate ratio is $\exp(0.8697) = 2.386$ and confidence intervals can be calculated using the error factor:

$$\exp(1.645 \times 0.3080) = 1.660.$$

The fact that the high energy-intake group is, on average, slightly younger than the low energy-intake group is the reason why the estimate of the rate ratio controlled for age is slightly smaller than the crude rate ratio (2.45). However, the difference is extremely small. This is not unusual; rather large differences between exposure groups in important variables are necessary for the effect of confounding to be appreciable.

Unfortunately it is not possible to calculate the values of M and S by hand using simple formulae. The computer programs which are used to carry out such computations are very flexible and allow more complicated models to be fitted. Accordingly discussion of these will be postponed until Part II and the remainder of this chapter will deal with methods which require only a hand calculator.

15.3 A nearly most likely value for θ

We saw in Chapter 13 that, in an unstratified analysis, both profile and conditional arguments led to the Bernoulli likelihood

$$D_1 \log(\Omega) - D \log(1 + \Omega),$$

where Ω, the odds for a case having been exposed, is $\theta Y_1/Y_0$. The gradient of the curve of log likelihood versus $\log(\theta)$ is

$$D_1 - D\frac{\Omega}{1+\Omega}$$

which, after substituting $\theta Y_1/Y_0$ for Ω and rearranging becomes

$$\frac{1}{Y_0+\theta Y_1}\left(D_1 Y_0 - \theta D_0 Y_1\right) = W\left(D_1 Y_0 - \theta D_0 Y_1\right),$$

where $W = 1/(Y_0+\theta Y_1)$. In a stratified analysis, the log likelihood is the sum of contributions of each stratum,

$$\sum\left[D_1^t \log(\Omega^t) - D^t \log(1+\Omega^t)\right]$$

and the gradient is similarly constructed by adding up gradient contributions:

$$\sum W^t\left(D_1^t Y_0^t - \theta D_0^t Y_1^t\right),$$

where $W^t = 1/(Y_0^t + \theta Y_1^t)$ are stratum *weights*.

The most likely value of θ occurs where the gradient is zero, that is, at

$$\theta = \frac{\sum W^t D_1^t Y_0^t}{\sum W^t D_0^t Y_1^t}.$$

Since calculation of the weights W^t involves θ, and this equation cannot be used directly to find the most likely value. However, it can be used *iteratively* as follows:

1. guess a value for θ, and use this to calculate initial weights;
2. using these, calculate a first estimate of θ;
3. using this new estimate, calculate more accurate weights.

The sequence of calculations may be repeated until there is no change in the estimate. Computer programs for maximum likelihood estimation use similar iterative methods of computation.

In practice, the estimate obtained is not very sensitive to changes in the values of the weights — rather large changes make only a relatively small difference to the estimate. Additionally, it may be argued that it is only really important to achieve the closest approximation to the log likelihood when estimating rate ratios which are fairly close to 1. These considerations suggest using the weights corresponding to the choice $\theta = 1$, and to go no further with the calculations. These weights are the reciprocal of the person-years observations in each age band:

$$W^t = \frac{1}{Y_0^t + Y_1^t} = \frac{1}{Y^t}.$$

Use of these weights leads to the *Mantel–Haenszel* estimate of the rate ratio*,

$$\frac{\sum D_1^t Y_0^t/Y^t}{\sum D_0^t Y_1^t/Y^t}.$$

In this expression, each age band makes contributions of

$$Q^t = \frac{D_1^t Y_0^t}{Y^t}, \qquad R^t = \frac{D_0^t Y_1^t}{Y^t}$$

to the top (numerator) and bottom (denominator) of the estimate respectively. The estimate of the rate ratio for age band t is Q^t/R^t and the combined estimate of the constant rate ratio is Q/R, where $Q = \sum Q^t$ and $R = \sum R^t$.

Exercise 15.1. Calculate Q^t and R^t for each of the three age bands in Table 15.1, and hence calculate the Mantel–Haenszel estimate of the rate ratio. Compare this with the most likely value.

15.4 Calculating p-values and confidence intervals

Approximate p-values are most easily calculated using the score test. Since the log likelihood for θ for the age-stratified comparison is the sum of contributions from each age band, it follows that its gradient, and hence the *score*, is the sum of scores for each stratum. Similarly, the curvature is the sum of the curvatures of the separate contribution of each stratum so that the overall score variance is the sum of score variances for each stratum. That is,

$$U = \sum U^t, \qquad V = \sum V^t.$$

Thus to carry out the test we first calculate scores and score variances for each stratum separately and then sum these over strata to obtain the total score and score variance. We then compare $(U)^2/V$ with the chi-squared distribution in the usual way. The contribution of stratum t to the score and score variance are of the same form as given at the end of section 13.2, namely

$$U^t = D_1^t - D^t\pi_\varnothing^t, \qquad V^t = D^t\pi_\varnothing^t(1-\pi_\varnothing^t),$$

where $\pi_\varnothing^t = Y_1^t/Y^t$, the ratio of exposed to total person years.

Exercise 15.2. For our example, what is the p-value for the null hypothesis that, after controlling for age, the rate ratio is 1.

*In fact Mantel and Haenszel did not propose this method but an extremely similar one for case-control studies. We shall discuss this in Chapter 18.

As before, the value of U may be interpreted as the difference between the number of cases who had been exposed and the number expected under the null hypothesis, taking into account the age structures of exposed and unexposed groups.

The calculation of the score variance, V, also allows us to calculate an approximate confidence interval around the Mantel–Haenszel estimate. A Gaussian approximation on the $\log(\theta)$ scale, with

$$S = \sqrt{\frac{V}{QR}}.$$

can be used to calculate an error factor and the approximate confidence interval in the usual way.[†]

Exercise 15.3. Calculate the standard deviation, S, of the log Mantel-Haenzsel estimate for the energy intake data. Use this to calculate a 90% confidence interval for the rate ratio adjusted for age.

These results are very close to those obtained using a computer program to find the Gaussian approximation to the log likelihood curve. The computer method is better in the sense that, as the quantity of data increases, the approximate interval of support approachs the correct likelihood-based interval, while the Mantel–Haenszel interval remains *slightly* wider no matter how much data we collect. The discrepancy is rarely important.

★ 15.5 The log-rank test

Our example in this chapter has involved stratification by a time scale, age, into three rather broad bands. In clinical follow-up studies time is measured from diagnosis or start of treatment and the incidence of events may vary rapidly, requiring the choice of narrow bands. This, together with the fact that choice of bands may introduce an arbitrary element into the analysis, has led to the popularity of a version of the test in which time is stratified infinitely finely into clicks, with no click containing any more than one event. This test is called the *log rank*[‡] or *Mantel–Cox* test.

Derivation of this test from that of the previous section is straightforward. The first thing to notice is that clicks which contain no event (i.e. with $D^t = 0$) make no contribution either to the score, U, or the score variance, V. We therefore need only consider those clicks in which we observe the occurrence of an event in one of the groups ($D^t = 1$). These are are

[†]This approximation is not widely known, but it would not appropriate to justify it here. It suffices to say that it is adequate for all our purposes.

[‡]This nomenclature may seem rather obscure, since the calculation of the test requires neither logarithms or ranks! It arises from an alternative derivation.

Table 15.2. Survival times in two groups of patients

Group	Time (days)
Test treatment ($N = 20$)	86, 99*, 119*, 123*, 139*, 161*, 185*, 212*, 231, 253*, 262*, 281*, 303*, 355*, 360*, 380*, 392, 467*, 499*, 514*
Control ($N = 20$)	73, 91, 102*, 120*, 135, 160*, 194, 202*, 209*, 220*, 252, 270*, 296, 330*, 347*, 375*, 390*, 414, 475*, 485*

known as *informative* time points.[§] Since each click is very short, we need not consider variation in the time spent by different subjects in the band, and the null probability that a failure was exposed becomes

$$\pi^t_{\oslash} = \frac{N^t_1}{N^t} = \frac{\text{Number of } \textit{exposed} \text{ subjects in study at time } t}{\textit{Total} \text{ number of subjects in study at time } t}.$$

Each failure makes a contribution to the score of the difference between the observed number of events in the exposed group, which is either 0 or 1, and the expected number, which is simply $\pi^t_{\oslash}$. The score variance is obtained by adding the contributions

$$V^t = \pi^t_{\oslash}(1 - \pi^t_{\oslash}).$$

Exercise 15.4. Table 15.2 shows times between entry to a clinical trial and relapse for patients receiving two methods of therapy. (The data are only illustrative — a real trial with so much censoring would need to be much larger than this!) The times marked with an asterisk represent times at which observation ceased without occurrence of relapse. Construct a table showing the times of occurrence of relapses, the number of patients in each group under study at each of these times, and the corresponding observed and expected relapses in the test group. Use this table to carry out the score test.

15.6 Comparison with reference rates: the SMR

An important special case concerns the comparison of age-specific rates in a study cohort, λ^t, with those in a *reference population*, which we shall denote by λ^t_{R}. We have discussed this informally in Chapter 6. A more formal treatment follows as a simple case of the methods discussed above.

The proportional hazards model holds that the ratio of age-specific rates in the study cohort to the reference rates is constant across age bands,

$$\frac{\lambda^t}{\lambda^t_{\mathrm{R}}} = \theta.$$

[§]Since clicks have no duration, we assume that no more than one event occurs at any time point.

If we observe D^t failures in Y^t person years of observation in each age band of the cohort, the log likelihood contribution is

$$D^t \log(\lambda^t) - \lambda^t Y^t$$

and making the substitution $\lambda^t = \theta\lambda_R^t$ this becomes

$$D^t \log(\theta) + D^t \log(\lambda_R^t) - \theta\lambda_R^t Y^t.$$

Since the reference rates λ_R^t are calculated from very large populations, they are effectively known constants, and the above log likelihood depends only on one unknown parameter, θ. The second term in the log likelihood does not depend on θ and can be ignored, and the third term may be simplified after noting that $\lambda_R^t Y^t$ is the expected number of failures obtained by multiplying the age-specific reference rate by the corresponding person-years of observation of the study cohort (see Chapter 6). Denoting this by E^t, the log likelihood contribution of one age band becomes

$$D^t \log(\theta) - \theta E^t$$

and summation over age bands leads to the total log likelihood

$$D \log(\theta) - \theta E,$$

where D, E are the total observed and expected numbers of failures. This is a Poisson log likelihood, but the rate ratio parameter θ replaces the rate parameter λ, and the expected number of failures E replaces the person-years Y. Thus estimating θ in this case is just the same as estimating a rate. The most likely value is the ratio of observed to expected cases, D/E, and in epidemiology this is called the standardized mortality ratio, or *SMR*. A 90% confidence interval can be calculated using the error factor

$$\exp\left(1.645\sqrt{\frac{1}{D}}\right).$$

An approximate p-value for the null hypothesis $\theta = 1$ can be carried out using the score and score variance

$$U = D - E, \qquad V = E.$$

Comparison of rates with reference rates in this way is known in epidemiology as *indirect standardization.*

Exercise 15.5. In the follow-up study of ankylosing spondylitis patients discussed in Chapter 6, the observed number of deaths from leukaemia was 31 while the expected number calculated from reference rates was 6.47. Calculate the 90% confidence interval for the common ratio of cohort age-specific rates to reference rates. Also calculate an approximate p-value for the null hypothesis $\theta = 1$.

Exercise 15.6. The calculation of the expected number of deaths in the ankylosing spondilitis study was based on person-years classified by both age and calendar period (see Chapter 6). What further modelling assumption is formally necessary to justify the analysis carried out in the previous exercise?

15.7 Comparing standardized rates

We showed in Chapter 14 that standardized rates estimate the marginal rates when the age distributions are corrected to a common standard. These are weighted sums of age-specific rates. In the case of three age bands, the marginal rate is

$$W^1\lambda^1 + W^2\lambda^2 + W^3\lambda^3$$

where (W^1, W^2, W^3) are the relative frequencies of the three age bands in the standard distribution, and the ratio of two marginal rates, corrected to the same age distribution, is

$$\frac{W^1\lambda_1^1 + W^2\lambda_1^2 + W^3\lambda_1^3}{W^1\lambda_0^1 + W^2\lambda_0^2 + W^3\lambda_0^3}.$$

When the proportional hazards model holds, every term in the numerator of this expression is θ times the corresponding term in the denominator, and it follows that the ratio of marginal rates will also be θ — the relationship between marginal rates is the same as that between the conditional (age-specific) rates. Thus, the ratio of standardized rates can be used as an estimate of θ. However it may not be a very good estimate if the standard age distribution gives high weight to age bands with few failures.

Note that the equivalence demonstrated above between the conditional and marginal comparisons does not hold for *all* stratification models. For example, if the ratio of the age-specific *odds* of failure for exposed and unexposed subjects is a constant, θ, for all ages then the ratio of marginal odds is not equal to θ, even when there is no confounding and the age distributions are identical. Thus we cannot always rely on the method of direct standardization if we are interested in comparisons within strata. In Chapter 18 we shall encounter an important example of this.

15.8 Comparison of SMRs

Although the ratio of standardized rates can be used as an alternative estimate of θ, there has been some controversy as to whether the ratio of two SMRs can also be used in this way.

An understanding of the formal model which lies behind indirect standardization clarifies this argument. Calculation of an SMR for an exposed cohort, using reference rates λ_R^t implies the model

$$\lambda_1^t = \theta_1 \lambda_R^t,$$

where θ_1 is the constant ratio of rates in this cohort to reference rates. Similarly, calculation of an SMR for an unexposed cohort implies the model

$$\lambda_0^t = \theta_0 \lambda_R^t.$$

A direct consequence of these two models is that the ratio of rates for the two cohorts is also constant across age. This can be demonstrated by simply dividing the two equations, when λ_R^t cancels leaving

$$\frac{\lambda_1^t}{\lambda_0^t} = \frac{\theta_1}{\theta_0} = \theta.$$

Thus if the age-specific rates for both exposed and unexposed cohorts are proportional to the reference rates, the comparison of SMRs is legitimate. Since the likelihoods for θ_1 and θ_0 are Poisson in form, with expected numbers of failures E_1 and E_0 replacing person-years observation Y_1 and Y_0, the likelihood for their ratio, θ, is the same as for the rate ratio in Chapter 13.

This method, however, relies on the assumption that both sets of age-specific rates are proportional to the reference rates. If they are proportional to each other, but not to the reference rates, then the ratio of SMRs will not correctly estimate the rate ratio θ. Because of this additional assumption concerning reference rates, estimation of θ by the ratio of SMRs is not usually to be recommended.

Solutions to the exercises

15.1 The calculations are as follows:

Age	Q^t	R^t
40–49	$2 \times 607.9/919.8 = 1.32$	$4 \times 311.9/919.8 = 1.36$
50–59	$12 \times 1272.1/2150.2 = 7.10$	$5 \times 878.1/2150.2 = 2.04$
60–69	$14 \times 888.9/1556.4 = 8.00$	$8 \times 667.5/1556.4 = 3.43$
Total	16.42	6.83

The Mantel–Haenszel estimate is $16.42/6.83 = 2.40$ while the most likely value is 2.39.

15.2 The score is:

$$U = \left(2 - 6\frac{311.9}{919.8}\right) + \left(12 - 17\frac{878.1}{2150.2}\right) + \left(14 - 22\frac{667.5}{1556.4}\right)$$

$$= 28 - 18.41$$
$$= 9.59$$

and the score variance is

$$V = 6 \times \frac{311.9 \times 607.9}{(919.8)^2} + 17 \times \frac{878.1 \times 1\,272.1}{(2\,150.2)^2} + 22 \times \frac{667.5 \times 888.9}{(1\,556.4)^2}$$
$$= 1.34 + 4.11 + 5.39$$
$$= 10.84.$$

The chi-squared value (1 degree of freedom) is $(9.59)^2/10.84 = 8.48$ and $p < 0.005$.

15.3 The standard deviation for the approximation is

$$S = \sqrt{\frac{V}{QR}} = \sqrt{\frac{10.84}{16.42 \times 6.83}} = 0.311.$$

The error factor for the 90% confidence interval is $\exp(1.645 \times 0.311) = 1.67$, and recalling that the Mantel–Haenszel estimate was 2.40, the confidence limits are $2.40/1.67 = 1.44$ (lower limit) and $2.40 \times 1.67 = 4.01$ (upper limit).

15.4 The times at which events occurred, the numbers of patients under observation, and the observed and expected relapses in the test group are shown below.

t	N_1^t	N_0^t	N^t	D_1^t	E_1^t
73	20	20	40	0	$20/40 = 0.50$
86	20	19	39	1	$20/39 = 0.51$
91	19	19	38	0	$19/38 = 0.50$
135	16	16	32	0	$16/32 = 0.50$
194	13	14	27	0	$13/27 = 0.48$
231	12	10	22	1	$12/22 = 0.55$
252	11	10	21	0	$11/21 = 0.52$
296	8	8	16	0	$8/16 = 0.50$
392	4	3	7	1	$4/7 = 0.57$
414	3	3	6	0	$3/6 = 0.50$

The overall score is

$$U = 3 - (.50 + .51 + .50 + \cdots + .57 + .59) = -2.13$$

and the score variance is

$$V = (.50 \times .50) + (.51 \times .49) + \cdots + (.50 \times .50) = 2.49.$$

The score test is $(U)^2/V = 1.82$ and $p > 0.10$. This test is the score test for $\theta = 1$ in the proportional hazards model which holds that the ratio of the relapse rates of the two treatments is constant (at θ) regardless of time since entry into the trial.

15.5 The most likely value of θ is the SMR,

$$\frac{31}{6.47} = 4.791.$$

The error factor is

$$\exp\left(1.645\sqrt{\frac{1}{31}}\right) = 1.344,$$

so that the 90% confidence interval is from $4.791/1.344 = 3.56$ to $4.791 \times 1.344 = 6.44$.
The score test is

$$\frac{(31 - 6.47)^2}{6.47} = 93.00$$

and $p < 0.001$.

15.6 Follow-up was stratified by both age and calendar period when calculating the expected number of deaths. The model which underlies the above analysis therefore assumes that the ratio of rates in the ankylosing spondilitis cohort to those in the reference population is constant for all ages and for all calendar periods.

15 Comparisons of rates within strata

15.1 The proportional hazards model

"***This problem is resolved by combining...***" [beginning of second paragraph]

Before we get to their modelling approach, it is appropriate to consider a simpler, more transparent way of combining the comparisons, namely by a **precision-based weighted average** of the observed stratum-specific rate ratios, or better still by precision-based weighted average of the *logs* of the stratum-specific rate ratios.

This too involves a model (a set of assumptions) in that it assumes (implicitly at least) that one is combing estimates of the same single but unknown parameter, $\log[\theta] = \log\left[\frac{\lambda_1^1}{\lambda_0^1}\right] = \log\left[\frac{\lambda_1^2}{\lambda_0^2}\right] = \log\left[\frac{\lambda_1^3}{\lambda_0^3}\right]$. In other words, it assumes that each $\log\,[\widehat{\lambda_1^t/\lambda_0^t}]$ is an estimate of the unknown scalar $\log[\theta]$, as if we had 3 estimates of *the* speed of light. This implicitly is the proportional hazards model.

Supplementary Exercise 15.1, using C&H's 3-age-strata example

i. Compute the logs of the 3 observed (empirical) stratum-specific rate ratios, and the variances of these.[1]

   Take a weighted average of these, using the inverses of the 3 variances as the weights.[2] Also, calculate the variance for this weighted average of logs.

   Finally, reverse scales by converting the point and interval estimate back to the θ, i.e., Rate Ratio, scale, and compare with the result at the top of column 2 of page 143.

ii. The reason for working in the log scale is that the distributions of the $\log[\hat{\lambda}]$'s (and their differences) are usually closer to Gaussian than the distributions of $\hat{\lambda}$'s are. This parameter transformation is covered in section 9.2.[3]

   Spiegelhalter in his book and in his work, is very keen on using 'Gaussian-looking' likelihoods – and Gaussian priors. Their 'conjugancy' in producing 'Gaussian-looking' posterior distributions simplifies matters considerably.

   But stratum-specific 'Gaussian-looking' likelihoods are also important for another reason, that is illustrated by this exercise.

   Consider the 3 age-strata in Table 15.1. and refer to Gaussian log likelihood in the first sentence of Chapter 9. In out case, their μ is our $\log[\theta]$. So lets pursue a Gaussian-likelihood-based estimate of μ.

   • Use the logs of the 3 observed log ratios calculated in part i above as your M_1, M_2 and M_3, (our 3 $\hat{\mu}$'s) and use the 3 variances calculated from part i as S_1^2, S_2^2, and S_3^2.

   • Plot the 3 separate log-likelihoods on the same graph, each with a different colour. (just by their widths, without the colours, can you identify which it which?)

   • Now plot the **sum** of the 3 log-likelihoods on this same graph, and measure where its reaches its maximum, and what the curvature is at this maximum.

   • Compare this calculated curvature with the sum of the inverses of the 3 quantities, S_1^2, S_2^2, and S_3^2. Comment.

   This exercise was intended to emphasize that it is easier to add Gaussian-based log-likelihoods if you write each one as $-(1/2)\,\tau\,(M-\theta)^2$, *where (as in the Bayesian framework),* $\tau = 1/\sigma^2$ *is referred to as the '**precision**.'*

Now to the author's model-based approach...

In this example, the p.h. model is used to reduce the dimension of the problem from a likelihood with $\underline{6}$ parameters for $\underline{6}$ rates

stratum(s) - 1	$\lambda_{s_1,1}$	$\lambda_{s_1,0}$
stratum(s) - 2	$\lambda_{s_2,1}$	$\lambda_{s_2,0}$
stratum(s) - 3	$\lambda_{s_3,1}$	$\lambda_{s_3,0}$

to one with $\underline{4}$ parameters for $\underline{6}$ rates

stratum(s) - 1	$\theta\lambda_{s_1,0}$	$\lambda_{s_1,0}$
stratum(s) - 2	$\theta\lambda_{s_2,0}$	$\lambda_{s_2,0}$
stratum(s) - 3	$\theta\lambda_{s_3,0}$	$\lambda_{s_3,0}$

[1] Remember the exercises from earlier chapter where you worked out the variance of the log of a rate ratio based on 2 Poisson r.v.'s D_1 and D_0 serving as numerators, and two know quantities Y_1 and Y_0 serving as the person-time denominators. Theoretically, it came out to $1/\mu_1 + 1/\mu_0$, so to be practical you need to plug in estimates of mu_1 and μ_0.

[2] It is a general result in statistics – and a commonly used question in exams to prove – that the 'best' (in the sense of minimum variance) linear combination of several estimates of the same parameter, where each estimate is accompanied by its own variance, is the one that uses the inverses of these variances as the weights.

[3] It turns out that, especially for parameters such as rates, the 1/3 power is also a very good 'Gaussianizing' transformation, but we will not pursue this further at this point.

and from there, by use of either a profile likelihood or a likelihood based on conditional distributions, to one with the $\underline{1}$ parameter of interest, θ, the (assumed constant over strata) rate ratio.

As we will see, in the case where the Poisson denominators are known (rather than estimates based on sampling), one can also fit the $\underline{4}$ parameter model by an $\underline{\text{un}}$conditional approach but focus only on the parameter of interest, θ.

15.2 "The" likelihood for θ

JH put quotes around The, since we need to be a bit more careful here. There are 3 possible likelihoods: 2 of them, the profile and 'conditional' likelihoods, are 1-dimensional, and happen in this case to coincide with each other, and the 3rd just-mentioned one, the 'unconditional' likelihood, which is 4-dimensional.

C&H use the 1-dimensional likelihood, and in particular the profile version.

They start with the 4-D log-likelihood

$$LL(\theta,\ \lambda_{s_1,0},\ \lambda_{s_2,0},\ \lambda_{s_3,0})$$

obtained as a sum of 6 cell-specific contributions. They then use profiling within each stratum to eliminate the $\lambda_{stratum,0}$ so that the 2 cells in a stratum contribute a stratum-specific $LL_{profile}(\theta)$ and so that the profile likelihood based on all 3 strata is

$$LL_{profile}(\theta) = \sum_i \{D_{s_i,1}\log(\Omega_i) - D_{s_i}\log(1+\Omega_i)\},$$

where, in stratum i,

$$\Omega_i = \theta \times (Y_{s_i,1}/Y_{s_i,0}).$$

Just as in chapter 13, with unstratified data, this profile log-likelihood is exactly the same as the log-likelihood for 3 binomial observations, each with its own Ω. However, all all 3 Ω's are connected by the single parameter of interest, θ, and three constants $Y_{s_1,1}/Y_{s_1,0},\ Y_{s_2,1}/Y_{s_2,0},\ and\ Y_{s_3,1}/Y_{s_3,0}$, so we can write this as a generalized linear model, with 3 binomial observations,

stratum	denom(D)	num(D_1)	$Y.ratio = \frac{Y_1}{Y_0}$	$\log(\Omega) = \log\left(\frac{E(D_1)}{D-E(D_1)}\right)$
1	6	2	0.513	$\log(\theta)+\log(0.513)$
2	17	12	0.690	$\log(\theta)+\log(0.690)$
3	22	14	0.751	$\log(\theta)+\log(0.751)$

Supplementary Exercise 15.2, based on C&H's 3-age-strata example

i. Create the function log-likelihood(θ) in `R`, and maximize it with respect to θ using `optimize` or otherwise. From the curvature, calculate the SE for $\hat{\theta}_{ML}$.

Repeat, but focusing on log-likelihood(β), where $\beta = \log[\theta]$, and compare your results with those of C&H.

Fig 15.1 shows both the 'exact' and Gaussian-based log-likelihood functions of $\log[\theta]$. Draw you own, to see if you get the same pattern.

ii. Did C&H pick a good example where a confounding factor (here age), if ignored, would lead to a very different (and very wrong) answer? Answer by calculating $\hat{\theta}_{ML}$ for the *aggregated* data (as in Chapter 13, i.e. before they segregated the data by age). Note the difference between aggregating *raw data* across strata, and aggregating *parameter estimates* (by summing likelihoods, or by some other weighting) across strata.

iii. Instead of explicitly defining and maximizing the conditional/profile likelihood, which simplifies in this example to a 1-parameter binomial-based likelihood, obtain $\hat{\theta}_{ML}$ using a GLM, for example, in `R`:

```
D1=c(2,12,14); D=c(6,17,22); X=c(1,1,1);
Y.ratio=c(311.9/607.9, 878.1/1272.1, 667.5/888.9);

fit=glm(cbind(D1,D-D1)~-1+X,family=binomial,offset=log(Y.ratio))
summary(fit)

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)
```

and verify that fitting this GLM leads to the same $\widehat{\log(\theta)} = 0.8697$ and $SE(\widehat{\log(\theta)}) = 0.3080$ that C&H report at the bottom/top of page 142/143.

iv. What would happen if you used the same p.h. model but fitted *all 4 parameters* in an *unconditional* approach? Use this code to see: you have 6 Poisson observations, the *link* is a log link, and the 6 $\log(y)$'s serve as *offsets*.

Comment on your results.

```
D.all6 = c(D1,D-D1) ; Index.category=c(1,1,1,0,0,0);
Y.all6=c(311.9,878.1,667.5, 607.9,1272.1,888.9);
Stratum=c(1:3,1:3)
cbind(Stratum,Index.category,D.all6,Y.all6)
Poisson.fit=glm(D.all6 ~ as.factor(Stratum) + Index.category ,
                family=poisson,offset=log(Y.all6))
summary(Poisson.fit)
exp(Poisson.fit$coefficients)
```

v. Remember to ask this term's MATH523 (GLM) teacher to explain to you why you get the (dis)agreement you get between the conditional and unconditional approaches to Poisson data, when in the unconditional approach you treat the stratum as a categorical variable.

[Optional] Also ask the teacher (or try it for yourself) whether you would you get the same odds ratio if you fitted an unconditional (2-binomial) model and the conditional (non-central hypergeometic) model to the frequencies {a=3,b=2,c=1,d=1}? *Hint*: see Breslow and Day, Volume I.

ML point- (& interval) Estimates via Newton-Raphson[4] Method

From `http://en.wikipedia.org/wiki/Newton's_method` ...

> In numerical analysis, Newton's method (also known as the Newton-Raphson method, named after Isaac Newton and Joseph Raphson) is perhaps the best known method for finding successively better approximations to the zeroes (or roots) of a real-valued function $f(x)$. Newton's method can often converge remarkably quickly, especially if the iteration begins "sufficiently near" the desired root. Just how near "sufficiently near" needs to be, and just how quickly "remarkably quickly" can be, depends on the problem. This is discussed in detail below. Unfortunately, when iteration begins far from the desired root, Newton's method can easily lead an unwary user astray with little warning. Thus, good implementations of the method embed it in a routine that also detects and perhaps overcomes possible convergence failures.
>
> Given a function $f(x)$ and its derivative $f'(x)$, we begin with a first guess x_0. A better approximation x_1 is
>
> $$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$
>
> Newton's method can also be used to find a minimum or maximum of such a function, by finding a zero in the function's first derivative.

In our case, we seek the root of the function $f(\theta) = dLogL/d\theta$, so the iteration takes the form

$$\hat{\theta}_{new} = \hat{\theta}_{previous} - \left. \frac{dLogL/d\theta}{d^2LogL/d\theta^2} \right|_{\theta=\hat{\theta}_{previous}}.$$

Exercises in previous years: Using the Newton-Raphson method, repeat Supplementary Exercise 3.1 (Estimation of σ^2 from grouped data), Exercise 3.2 (Estimation of concentration via a dilution series) via the Newton-Raphson method; Supplementary Exercise 5.1 (Estimation of (constant across time-bands) rate parameter λ from censored HIV data).

[4] JH included this NewtonRaphson technique in the early years of BIOS601, before he was introduced to the `optimize` and `optim` functions in `R`. He still believes students should know this technique and be able to 'roll their own' maximization routines when needed.

15.3 $\theta = RateRatio \rightarrow \hat{\theta}_{Mantel-Haenszel} \simeq \hat{\theta}_{ML}$ *a l m o s t* !

The key is the form of the θ estimator shown in the middle of page 144.

$$\hat{\theta}_{ML} = \frac{\sum D_{s_i,1} \times Y_{s_i,0}/(Y_{s_i,0} + \hat{\theta}_{ML} Y_{s_i,1})}{\sum D_{s_i,1} \times Y_{s_i,0}/(Y_{s_i,0} + \hat{\theta}_{ML} Y_{s_i,1})}$$

C&H note that the profile and conditional likelihoods are both the same, and are based on the fitting of 3 binomials with different Ω's, as above. You can work through their math at the top of page 144. You can also arrive at the estimator by using an ***estimating equation*** directly. In this case, we are estimating a single parameter θ and so there is only 1 estimating equation, and as in all generalized models, the first estimating equation is that the sum of the observed y values must equal the sum of the *fitted* or *expected* values. In our case, the 3 observed values are $D_{s_1,1},\ D_{s_2,1}$ *and* $D_{s_3,1}$, so the estimating equation is

$$\sum_i D_{s_i,1} = \sum_i E[D_{s_i,1}] = \sum_i \widehat{D_{s_i,1}}.$$

Now

$$\widehat{D_{s_i,1}} = D_{s_i} \times \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1},$$

and so our estimating equation is

$$\sum_i D_{s_i,1} = \sum_i D_{s_i} \times \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1}.$$

If we now break up each D_{s_i} into its two components, we get

$$\sum_i D_{s_i,1} = \sum_i D_{s_i,0} \times \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1} + \sum_i D_{s_i,1} \times \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1}.$$

After re-arranging terms, we get

$$\sum_i D_{s_i,1} \left(1 - \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1}\right) = \sum_i D_{s_i,0} \times \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1},$$

or

$$\sum_i D_{s_i,1} \left(\frac{Y_0}{Y_0 + \hat{\theta} Y_1}\right) = \sum_i D_{s_i,0} \times \frac{\hat{\theta} Y_1}{Y_0 + \hat{\theta} Y_1},$$

or as

$$\hat{\theta} = \frac{\sum_i \frac{D_{s_i,1} \times Y_0}{Y_0 + \hat{\theta} Y_1}}{\sum_i \frac{D_{s_i,0} \times Y_1}{Y_0 + \hat{\theta} Y_1}},$$

or, with $W_i = 1/(Y_0 + \hat{\theta} Y_1)$, as

$$\hat{\theta} = \frac{\sum_i W_i \times D_{s_i,1} \times Y_0}{\sum_i W_i \times D_{s_i,0} \times Y_1}.$$

One can also arrive at this as C&H did on p 144, by setting to zero the sums of the 3 derivatives of the profile log likelihood with respect to $\log(\theta)$, or with respect to θ itself, and finding the root. Either way, the estimating equation is always the 'balancing equation',

$$\sum_{strata} observed\ no.\ of\ exposed\ cases = \sum_{strata} fitted\ no.\ of\ exposed\ cases,$$

used above.

Supplementary Exercise 15.3 Follow the iterative re-weighting scheme 1.2.3. described by C&H on the bottom of page 144 to arrive at $\hat{\theta}_{ML}$.

Note: This particular iterative re-weighting scheme produces the ML point estimate, but does not provide a measure of precision for it. The Newton-Raphson procedure and the optimize procedure do, since they use an analytical or numerical version of the second derivative of the log-likelihood.

Note also that if we write the log-likelihood as a function of $\beta = \log(\theta)$, rather than θ itself, and carry out the N-R (or `optimize`) procedure on the β scale to obtain a ML point estimate of β, then to get back to the CI on the Rate Ratio scale, you would use the SE on the log scale to get a symmetric (z-based) CI for β, then convert it a CI for $\theta = \exp(\beta)$

[**For the curious**] The genius behind the M-H method is its stability. Although it can be algebraically re-written as a weighed sum of ratios (a dangerous thing to do), such a re-expression goes against the very spirit of the estimator: it is meant to be a *single* ratio of two sums, a 'numerator' sum, and a 'denominator' sum. Miettinen, in his interview with JH, tells of his conversation with Mantel, and Mantel's explanation of how he came up with what C&H calls the weights $W = 1/(Y_0 + \theta Y_1)$. Mantel said that if he used as a numerator the sum of products of the form $D_1 Y_0$, and as a denominator the sum of products of the form $D_0 Y_1$, these individual products would be too volatile and 'jumpy' , and would 'wobble' by far more than their information content justified. So he decided to divide each product by $Y_0 + Y_1$ so as to 'slap down' the products, and not have them vary so much.

Exercise [Optional] Simulate the variability of a M-H-type estimator that does not 'slap down' or 'tame' the products, i.e. an estimator of the form

$$Q = \sum D_1 Y_0;\ R = \sum D_0 Y_1; \quad \hat{\theta} = \frac{Q}{R}.$$

Supplementary Exercise 15.4 See section (c) of section 3.6 of Breslow and Day Vol II.[5] There they say that the ML estimation of the rate ratio ψ from stratified PT data requires iterative calculations, so let's iterate...

We will use B&D's Example 3.11, with data, shown in Table 3.14, from $J = 13$ age-period strata.

Table 3.14 Series of 2 × 2 tables used in example 3.11. Low exposure (−) means less than 1 year of heavy or moderate arsenic exposure; high exposure (+) means 15+ years

Age (years)		Calendar period							
		1938–1949		1950–1959		1960–1969		1970–1977	
	Exposure	−	+	−	+				
40–49	$d/\hat{d}$	2/1.50	0/0.50	0/0.00	0/0.00				
	n	3075.27	337.29	936.75	121.00				
	$\hat{\psi}$	0.00		—					
	Exposure	−	+	−	+	−	+		
50–59	$d/\hat{d}$	2/3.58	4/2.42	3/4.02	3/1.98	3/2.52	1/1.48		
	n	2849.76	626.72	2195.59	349.53	747.77	142.33		
	$\hat{\psi}$	9.0		6.3		1.8			
	Exposure	−	+	−	+	−	+	−	+
60–69	$d/\hat{d}$	2/5.52	9/5.48	7/7.73	7/6.27	10/8.65	3/4.35	1/1.17	1/0.83
	n	2085.43	672.09	1675.91	441.10	1501.73	244.82	440.21	100.64
	$\hat{\psi}$	14.0		3.8		1.8		4.4	
	Exposure	−	+	−	+	−	+	−	+
70–79	$d/\hat{d}$	3/1.98	1/2.02	6/4.32	2/3.68	6/4.40	1/2.60	6/5.62	2/2.38
	n	833.61	277.25	973.32	268.27	1027.12	197.20	674.44	92.75
	$\hat{\psi}$	1.0		1.2		0.9		2.4	

d = observed deaths; $\hat{d}$ = fitted deaths under ML estimate of common rate ratio; n = person-years denominator; $\hat{\psi}$ = rate ratio in each table

Again interest is in the rate ratio parameter $\psi = \lambda_{j1}/\lambda_{j0}$, assumed (for now) to be constant over the $J = 13$ strata.

Thus, for each of the J strata, $O_{j1} \mid D_j \sim Binomial(D_j, \pi_j)$, where

$$\pi_j = \psi \times PT_{j1}/(\psi \times PT_{j1} + PT_{j0}).$$

Note the switch of notation, from O_{j+} to D_j, and subscripts 1 and 0 for exposed and not.

i. Derive the ML estimating equation (3.15) for $\hat{\psi}_{condn'l}$,

$$\sum_{j=1}^{j=J} D_{j1} = O_1 = E_1 = E\left(\sum_{j=1}^{j=J} D_{j1}; \psi\right) = \sum_{j=1}^{j=J} D_j \psi n_{j1}/(n_{j0} + \psi n_{j1}),$$

by obtaining the expression for $d \log L/d\psi$ and setting it to zero.

B&D say that

> *In large samples the most accurate estimator of ψ is the maximum likelihood estimate, obtained by setting the overall observed number of deaths D_2 in the exposed group (index category) overall equal to its expected value.*

ii. Use the Newton-Raphson iterative method to find the root of the $d \log L/d\psi$ function, ie

$$\hat{\psi}^{(k+1)} = \hat{\psi}^{(k)} + \frac{d \log L/d\psi}{d^2 \log L/d\psi^2}\bigg|_{\hat{\psi}^{(k)}} = \hat{\psi}^{(k)} + \frac{\Sigma_j d \log L_j/d\psi}{\Sigma_j d^2 \log L_j/d\psi^2}\bigg|_{\hat{\psi}^{(k)}}.$$

iii. How does the iteration change if we rewrite the Likelihood, and thus the log Likelihood, in terms of β, where $\psi = \exp(\beta)$?

iv. Obtain $\hat{\psi}_{condn'l}$ from a generalized linear model (Binomial) fitted to the 13 binomial observations. Note that one can specify Binomial (rather than Bernoulli) data by using as 'y' a matrix with 2 columns: the numbers positive and negative, i.e.

`glm(cbind('# +ve' vector,'#no. -ve' vector) ~ ..., family=binomial, ...).`

v. Obtain $\hat{\psi}_{uncondn'l}$ from a generalized linear model (Poisson, 14 parameters) fitted to the $(j = 1, \ldots, 13) \times (i = 0, 1) = 26$ observations $\{O_{ji}, PT_{ji}\}$.

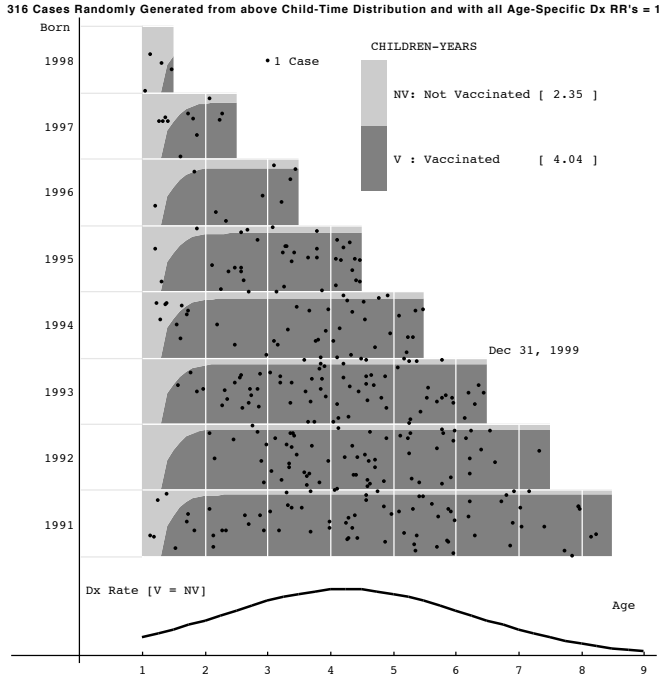
Are your estimates in agreement with Breslow and Day's statement (lines 5-6, page 109) that under the Poisson model, $\hat{\psi}_{condn'l} = \hat{\psi}_{uncondn'l}$?

Note B&D's comment that the same will **not** be true for conditional vs. unconditional estimation of a common rate ratio **when the PT's are estimated** from J stratified denominator ('control') series, particularly if the strata are sparse.

[5] The various chapters can be found in the link to bios602-2009 in the Resources for chapter 15. The 2 volume of Breslow and Day's books (Vol I: case-control studies; Vol II: cohort studies) are also now downloadable as .pdf files from `http://www.iarc.fr/en/publications/pdfs-online/stat/`

Supplementary Exercise 15.5: Is there a higher rate of autism in children who have been vaccinated with MMR? And, does it matter whether we correct/adjust for age?

Autism cases are shown as dots, and the vaccinated and unvaccinated child-time as darker and lighter areas. Notice that the vaccinated child-years occur at younger ages (average 2.35 years), where (as is shown at bottom) the rates of autism diagnosis are lower – so a simple (age-blind) comparison of the autism density in darker and lighter areas (average ages 2.35 and 4.04 years) would also be a comparison of rates in older versus younger years, and so reflect a mix of the effect of age and the effect of vaccination. [6].



The locations of the 316 cases in this modification of the Lexis diagram were randomly generated by ...

1 Calculating the "rate of diagnosis by age" curve (arbitrary scale) at ages=1.25 to 8.25 in steps of 0.5 (i.e. at 15 age-points; to simplify your job of counting cases in the various age cells, the diagram shows coarser, 1 year , i.e., birthday, boundaries)

2 Multiplying these "rates" by the numbers of children "in view" at each of these that ages, to get, for each of the 15 vertical age-slices of "child-time", a number proportional to the expected number of cases in that vertical child-time slice; then scaling the 15 expected numbers summing to 316.0: expect an average of 19.0 to be diagnosed between 1 and 1.5 years of age, 23.5 b/w ages 1.5 and 2, ... 31.1, 33.2, 38.8, 35.5, 36.6, 28.4, 25.9, 16.6, 13.3, 6.71, 4.76, 1.58, ... 0.992 between ages 8 and 8.5.

3 For each age-slice, randomly generating a count from a Poisson distribution with the corresponding expected value. Repeat until the sum of the observed number of cases is in fact 316, as it was in the actual study. This gave 19 between 1 and 1.5 years of age, 19 between ages 1.5 and 2, and so on, .. 23, 27, 37, 35, 42, 31, 27, 24, 13, 7, 5, 5, ... 2 between ages 8 and 8.5.

4 For each of these cases, randomly choose a year of birth (i.e. randomly along the vertical scale, *without regard to whether the location will be in a unvaccinated or a vaccinated child-time cell.*) and a more refined age at diagnosis (randomly within the 0.25 age-band on each side of 1.25, or 1.75, or etc. *,without regard to light/dark*). If the random location is in the darker(lighter) area, the case involves a child who was (un)vaccinated at the time of diagnosis.

EXERCISE: From the diagram, (manually) count the vaccinated and unvaccinated cases (numerators) in each vertical age-slice. Estimate (roughly) the (relative) sizes of the corresponding vaccinated and unvaccinated child-years (denominators) [hint: the proportions vaccinated by the end of the study range from 0.92 (1991 cohort) to 0.88 (1994), to 0.84 (1997), to 0.55 (1998)]. Using these numerators and denominators, calculate *an age-adjusted* RR.

i. On the website you will find R code to read the data into a data frame with 72 records: 2 'exposure' levels (vaccinated/un-vaccinated) × 36 cells. The experience inside each cell is from the same Lexis square or rectangle, where the child years come from children in a single-age and single year 'bin' or 'rectangle'.

   Analyze the data using (unconditional) Poisson regression, as the authors did, using a 36-level variate for 'cell' and a binary indicator (dummy) variate for the 'vaccinated' category (1='yes'; 0-'no'). Don't forget to include the (36) offsets [see the simple 3-strata example above]

ii. You can also use the R code provided later in the same file to set up the data for the binomial-based analysis: 36 binomial observations, each with its own offset, 1 per age-year cell.

   Analyze the data using conditional Poisson regression, i.e. using a binomial model, a binary indicator (dummy) variate for the 'vaccinated' category (1='yes'; 0-'no'), Don't forget to include the (36) offsets [see the simple 3-strata binomial example above]

iii. Now use the same 36-row data frame to calculate $RateRatio_{M-H}$, i.e., the 'almost MLE' (C&H's name for the Mantel-Haenszel-type) Rate Ratio estimate.

iv. How close are the 3 estimates? Does it matter in this example that we adjusted for age? (answer by comparing them with the 'crude' RateRatio, which you will have already computed in an earlier exercise).

[6] Just like the confusion in the case of the Belfast Catholic girl & Protestant boy.

Supplementary Exercise 15.6: Do Oscar Winners Live Longer than Less Successful Peers? A Reanalysis of the Evidence

The aims are to carry out (1) the 'P-Y' analysis described in the 2006 'McGill' re-analysis, and (2) calculate the 'fewer-assumptions involved' Mantel-Haenszel summary ID ratio that the McGill authors calculated but – not to confuse the reader with yet another analysis – omitted from the article. Later on in the course, we will analyze the data with the same (time-dependent Cox PH) model that was reported on in the 2006 article.

Under the EPIB634 Resources for regression models for (incidence) rates, you will find (a) the Oscar data set[7] with one data-record per performer (b) a dataset (with approx. 20,000 records) in which each the performer's data-record has been converted (split) into 1-year data-records, and classified according to age, period, AND Oscar-status, (c) a smaller dataset in which the individual performer-years (and numbers of deaths) have been aggregated into 'sex-age-period-Oscar' cells, with 5-year age-bands and 10 year calendar-year-bands,[8] and (d) a file similar to (c), but where *all* of a performer's performer-time is allocated to the 'winners' category if that performer *ever* won an Oscar, or to the 'nominated' category if (s)he was nominated but never won.[9]

In the *description* of (b) and (c) below, the name of the Oscar-status indicator is shortened to O, with $O = 0$ indicating performer-time lived as a nominee, and $O = 1$ indicating performer-time lived as an Oscar winner. In the *actual dataset to be analyzed, i.e. in* (c), $O = 0$ corresponds to `w.cat=0` and $O = 1$ to `w.cat=1`.

In (b) each (Oscar-status-specific) record documents the experience in each (age, period) 'rectangle'[10] traversed, i.e., the number of years spent in that rectangle , and the Vital status (0 if alive, 1 if dead) at the end of these years.[11] Because the Lexis program is written for generic *transitions ('events') of any type (not necessarily bad ones)*, this status variable is called `lex.Xst`, which refers to the status (in our example *vital* status, 0 alive, 1 dead) at the performer's 'exit' (pardon the pun, but the '`X`' in '`Xst`' stands for an *epidemiologic* 'exit' from the Lexis diagram, and the '`st`' stands for status). The other key variable is `lex.dur`, which refers to the `dur`ation or length of the performer's time-slice.

In (c), which is formed by summing the performer-time `lex.dur` and the `lex.Xst` over all transits through the same sex-age-period-O cell, the two sums are the *total p-t* and *total deaths* in this cell – remember that a sum of 0's and 1's is a count of the number of 1's.

i. Use dataset version (c) to compare the death rates in the performer-years lived as nominees (reference category, `w.cat=0`) with those lived as winners (index category, `w.cat=1`), by fitting the following multiplicative (i.e. 'rate ratio') model[12] to the numbers of deaths in each sex-age-period-Oscar (shortened to s-a-p-O here, in order to fit the equation into one line) 'cell'.

$$Rate_{cell} = Rate_{ref.cell} \times M_{s:ref} \times M_{a:ref} \times M_{p:ref} \times M_{O:ref},$$

where the $ref.cell$ is a suitably chosen reference 'corner' cell (Clayton and Hills' terminology), and each M (the rate 'Multiplier') is short for Mortality Rate Ratio (MRR), – the theoretical, unknown, to be estimated, ratio of the mortality rate in the category[13] of the determinant in question relative to the reference category of that determinant.

For fitting purposes, you translate the *epidemiologic* (rate) model above into the following *statistical* model

$$E[\#deaths] = e^{\{logRate_{ref} + logM_s \times s + logM_a \times a + logM_p \times p + logM_O \times O + \log(PT)\}},$$

[7] For reasons JH can better explain in person, this differs slightly from that analyzed in the Redelmeier article.

[8] You are asked to the analyses with (c), which is named `aggregated-Lexis-rectangles.txt`. Nowadays, with fast computers and lots of live memory / disk storage space for large datasets, you *could* do the analysis using (b). Since it uses finer subdivisions of age and calendar period, you would get get slightly different answers, and you would probably choose to model age and calendar-time with (functions of) continuous variables, rather than with a very large number of indicator variables – 'dummy' variables, if you insist on that meaningless term – for the finer age- and calendar-period categories.

[9] The name of datafile (d), `aggregated-Lexis-rectangles-r.txt`, has the suffix '`-r`' to denote it as the 'Redelmeier' allocation of the performer-time.

[10] This terminology is from Lexis, who tended to use squares, e.g., 5-year age bands and 5-year calendar-year bands: since death rates vary faster over ages than over calendar time, you want to make the age-bands (i.e., the age-matching) quite narrow: thus jh formed rectangles that are 1 (age) year high by 10 (calendar) years wide, so in effect each slice was 1 year long: you could rerun the time-slicing program with other 'cuts.'

[11] If you want to see how these split records were created, you can look at and run the `R` code shown in the resources. It uses the `Lexis` package that is available from the `R` site, and developed by Carstensen (`R` 'Epi' package `http://staff.pubhealth.ku.dk/~bxc/Epi/`). See also the `survSplit` function in the `survival` package – we used this to split the time in the COMPARE (stents) study. One of the students in bios602 discovered two other options. One is a standalone Windows program, from `http://epi.klinikum.uni-muenster.de/pamcomp/pamcomp.html`; the other is the `pyears` function in the Survival package in R (jh doesn't remember if Survival is part of the default R installation, or needs to be added). **Stata** users: there is a time-slicing function used in conjunction with survival analyses.

[12] One could, and would if need be, refine this model further, e.g. by refining the relationship of rates with age, and allowing for the possibility of different effects of O in males and females...

[13] Or *level*, if we model the variable as an interval variable.

so that

$$\log\{E[\#deaths]\} = \beta_{ref} + \beta_s \times s + \beta_a \times a + \beta_p \times p + \beta_O \times O + \log(PT).$$

Writing out both models lets you match the coefficients from the fitted statistical (`R`) model with the fitted parameter value(s) of interest in the epidemiological (rate) model. (def'n.: *epidemiologist*: a student of *rates*).

ii. Write out the fitted multiplicative model in the same way as Clayton and Hills did in Table 22.7 in their Introduction to Regression chapter of their Statistical Models for Epidemiology textbook. Comment on the MRR for the 'years lived as a winner' vs. 'years lived as a nominee' contrast.

iii. Comment on the fitted effects of gender[14], age and calendar time, and whether they 'fit' with what you expect, and have seen in other datasets.[15]

iv. From dataset (c) calculate the total performer-time lived as a nominee ('$PT_{nominee}$'), and the total performer-time lived as a winner ('PT_{winner}'). Compare these with the corresponding values calculated from the 'Redelmeier' version, i.e., from dataset (d). Comment.[16]

v. Fit the same multiplicative model fitted in (i) to the data in dataset (d). Compare the fitted 'O' effect in this dataset – where `w.cat` is a fixed-from-the-outset variable – with what you found in the (McGill) version – where `w.cat` is a time-dependent variable. Comment.

vi. How would Mantel have analyzed these data? The R code file in resources includes some that allows you to convert datafile (c) into a form where you can treat sex, age and calendar period as stratifying variables – it puts the 'exposed' PT and deaths in the exposed PT in the same data-record as those for the un-exposed PT in the same stratum, making it easy to obtain the stratum-specific products, and to obtain the numerator and denominator sums used to calculate the ratio in formula 8.5 – déjà vu – in Rothman2002.

Use this re-arranged dataset to calculate this Mantel-Haenszel mortality rate ratio. How does it compare with the one obtained from Poisson regression?

vii. Use this same re-arranged dataset to calculate separate Mantel-Haenszel mortality rate ratios for actors and actresses. Based just on the numbers of deaths involved, do you think they are statistically significantly different?

If you wanted to pursue this effect-modification numerically, you could use the formula to obtain the SE of each rate ratio (or rather the SE of the log-rate-ratio). The formula is given in section 3.6(d) of Breslow and Day Volume II. It is quite tedious to do by hand, but quite easy with `R` or Excel.

viii. Use this same re-arranged dataset to obtain a 'MLE' from the profile (or conditional) 1-parameter likelihood – i.e. 'profile-out' or 'condition-out' all of the other parameters in the unconditional model you fitted in part *i*, so the focus is just on the rate ratio for the index vs. reference categories of the determinant of prime interest. *Hint: this problem has the same structuer as the one in supplementary exercise 15.7.*

The original report continues to be cited... just Google 'Oscars longevity'

`http://www.health.harvard.edu/press_releases/oscar_winners`

15.4 P-values from Score Test

Supplementary Exercise 15.7

Using the computations for C&H's exercise 15.2, based on their 3-strata example, as a template, carry out the corresponding calculations for the Autism data – using the broader (1-year wide) age-strata. Repeat with the narrower age-strata.

15.5 The log-rank test

The introduction of this test at this place in this chapter is a bit unusual, as none of other examples in this chapter as arises from a trial, and there is no natural 'time-zero.' His main point, and Mantel's point when he saw the log-rank test introduced, was that it is a test that was already in use in a 'stratified data' context in the comparison of rates in epidemiology. Thus, if

[14] Even though we used the term 'sex' above, one could make a good argument for preferring the term 'gender' in this context: Google 'gender vs. sex'.

[15] The effects of gender, age and calendar time are secondary here, but if you do choose to represent age and calendar-time as linear (continuous) variables, make sure you report their effects correctly – they should broadly 'line up' with the fitted effects when using indicator variables.

[16] For the principle behind the correct allocation of person-time, and early examples of incorrect P-T allocation, see section 3.1 of Volume II of Breslow and Day's text, available in the resources for the bios602 course. See also the material on 'immortal-time' bias in the 'Regression models for (incidence) rates' resources on the 634 website.

one thinks of each 'risk set' as a narrow time-slice or time-stratum, then one can see how indeed the M-H approach applies.

The M-H test is of course much broader, and (as is obvious in the example from Breslow and Day, and in M-H's original 1958 illustration) the strata do not have to be *time*-based.

See elsewhere for JH's illustration of the log-rank test (and explanation of how the ranks come into the name of the test) in his comparison of the longevity of the Titanic survivors with that of the general population.

Supplementary Exercise 15.8

Consult some textbooks or Internet web pages to see how the log-rank test is usually presented (JH's own course webpages have some examples).

You will notice that some of them focus *only* on the observed and expected-under-the-null numbers of events in the *index* category (effectively the 'a' cell in each 2×2 table, and on the sum (over strata or risk sets) of their differences, squaring this sum at the end, and comparing it with the sum of the null-variances of the individual 'a-E$[a|\text{Null}]$' statistics. This approach is very much in the same spirit as the test statistics introduced by M&H, and the test derived by summing the Scores.

The other approach is a bit more like the traditional chi-square test for a single 2×2 table, but where the summation is over the *two* contrasted categories, rather than using just the 'a' cell.

i. List one source for each of these two approaches. Do you think they would lead to very different P-values in any practical situations? Include one example, worked both ways.

ii. Consider a single 2×2 table, with 'a' representing the frequency in (say) the upper left cell. Show that the arithmetic in the 'usual' $\sum(O-E)^2/E$ form, with the $\sum$ taken over all 4 (a, b, c, d) cells, leads to the same null chi-square statistic as the form

$$\frac{(a - \text{E}[a|\text{H}_0)^2}{Var[a - \text{E}[a|\text{H}_0]}.$$

List any assumptions you had to make to get the algebra to work out. You may want to consult JH's sections 2 and 3.3 on chi-square tests in his notes on 'Analysis of Proportions via Chi-Squared tests' in the 2007 version of bios601. You can also look at the portion on 2×1 tables in section 4.

iii. Before doing this next sub-question, ask three people who have taken a survival analysis course [or consult three textbooks or online course notes or blogs or websites] *why is log-rank-test called the log-rank-test?* and report their responses.

iv. Read the presentation 'A finely stratified log-rank test with effectively-infinite-size comparison groups' which you can find by searching within JH's course webpages [his home page has a search box].

Having done so, how you would now respond if that same question were directed at you?