Short Articles
International Journal of Epidemiology
International Epidemiological Association Inc. 1984
Vol. 13, No. 1
Printed in Great Britain

Elementary Methods of Cohort Analysis*

HANLE.

N E BRESLOW

Breslow N E (Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195, USA). Elementary methods of cohort analysis. *International Journal of Epidemiology* 1984, **13**: 112–115.
The Mantel–Haenszel procedure offers a simple and efficient means of estimating a common rate ratio from incidence density data in cohort studies. A new formula is provided for the variance of its logarithm, comparisons are made with the method of maximum likelihood, and associated tests for heterogeneity and trend in the component rate ratios are described.

A common problem in cohort analysis is the estimation of a summary incidence or mortality rate ratio for exposed versus unexposed persons while adjusting for the effects of confounding variables by stratification of the sample. Observations are typically arranged in 2×2 tables showing numbers of cases or deaths and person-years denominators in each stratum (Table 1). As an example, the left hand columns of Table 2 present data for coronary deaths among smokers and non-smokers from the British doctors' study.[1] Rothman and Boice,[2] subsequently denoted R&B, use these same data to illustrate statistical techniques that they have programmed for hand held calculators. The present article reviews their methods and suggests some additions so as to provide a coherent and comprehensive set of tools for cohort analysis.

THE STATISTICAL MODEL

An accurate approximation to the sampling distribution of the data in Table 1 is to assume that the numbers of deaths d_{1i} and d_{2i} in the ith of I strata follow independent Poisson distributions with means $\lambda_{1i} n_{1i}$ and $\lambda_{2i} n_{2i}$, where λ_{1i} and λ_{2i} are the unknown disease incidence or death rates. The key parameters are the rate ratios $\psi_i = \lambda_{1i}/\lambda_{2i}$ for exposed versus unexposed. Several hypotheses of interest are:

H_0: $\psi_i = 1$, the global null hypothesis;
H_1: $\psi_i = \psi$, the hypothesis of a common rate ratio;
H_2: $\psi_i = \psi \cdot f(\theta x_i)$, the alternative of trend;
H_3: ψ_i unrestricted, the general alternative.

In H_2 f denotes any smooth increasing function such that both f and its first derivative take the value 1 at $\theta x = 0$, e.g., $f(\theta x) = 1 + \theta x$ for a linear relationship or $f(\theta x) = \exp(\theta x)$ for a log-linear one. This hypothesis assumes that there are quantitative variables x_i associated with each of the I ordered strata, and it may not be appropriate in all applications. When it is, one often simply sets $x_i = i$ to test for a trend in the rate ratios with age or other ordered variables.

The usual goal of the statistical analysis is to test the null hypothesis, estimate the rate ratio assuming it is common to all strata, and evaluate this latter hypothesis relative to alternatives of trend or heterogeneity. General principles of inference[3] suggest that one consider a distribution for the data that depends only on the parameters of interest. This is easily accomplished here since the d_{1i}, conditional on the total deaths $D_i = d_{1i} + d_{2i}$ in each stratum, are binomial with denominators D_i and probabilities $p_i = \psi_i n_{1i}/(\psi_i n_{1i} + n_{2i})$.

TEST OF THE NULL HYPOTHESIS

The efficient score test[3,4] of H_0 versus H_1 based on the model simply compares the total number of deaths among the exposed to that expected if the rates for exposed and unexposed were equal within each stratum. It is a variant of the classical Cochran–Mantel–Haenszel[5,6] test whose initial use for cohort

TABLE 1 *Data layout for rate ratio estimation in a cohort study.*

Stratum		Exposed	Unexposed	
i	Deaths	d_{1i}	d_{2i}	D_i
	Person-years	n_{1i}	n_{2i}	N_i

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA and German Cancer Research Center, Heidelberg, West Germany.

TABLE 2 *Deaths from coronary disease among British male doctors.*[*]

Age	Person-years		Observed deaths		Expected deaths[†]		Rate
	Smokers	Non-smokers	Smokers	Non-smokers	Smokers	Non-smokers	ratio
i	n_{1i}	n_{2i}	d_{1i}	d_{2i}	$D_i\hat{p}_i$	$D_i\hat{q}_i$	
35–44	52 407	18 790	32	2	27.17	6.83	5.73
45–54	43 248	10 673	104	12	98.88	17.12	2.14
55–64	28 612	5710	206	28	205.26	28.74	1.47
65–74	12 663	2585	186	28	187.19	26.81	1.36
75–84	5317	1462	102	31	111.49	21.51	0.90
Totals	142 247	39 220	630	101	630.00	101.00	1.72

[*] Data from Doll and Hill[1] as quoted by Rothman and Boice.[2]

[†] Estimated by maximum likelihood under the hypothesis of a common rate ratio.

analysis is ascribed by R&B to Shore et al.[7] The formula is

$$T = \frac{\Sigma(d_{1i} - D_i\, n_{1i}/N_i)}{\{\Sigma d_i\, n_{1i}\, n_{2i}/N_i^2\}^{1/2}} \qquad (1)$$

where here as elsewhere Σ denotes summation over $i = 1, 2, \ldots, I$. For the data in Table 2 we find $T = 3.32$ and, referring T to tables of the normal distribution, a two-sided p-value of 0.001.

MAXIMUM LIKELIHOOD ESTIMATE AND VARIANCE

The maximum likelihood estimate of the common rate ratio,[8] which we denote $\hat{\psi}_{ML}$, is obtained by equating the observed number of deaths among the exposed to that expected under H_1:

$$\Sigma d_{1i} = \Sigma D_i\, \psi n_{1i}/(\psi n_{1i} + n_{2i}). \qquad (2)$$

Solution of this equation requires iteration but is programmed by R&B or available with GLIM[9] or other standard programs. For a GLIM analysis, we note that the probability p_i that a death in the ith stratum was exposed may be written

$$\text{logit}(p_i) = \log\{p_i/(1 - p_i)\} = \log(n_{1i}/n_{2i}) + \log(\psi).$$

H_1 thus defines a linear logistic model in which the known quantities $\log(n_{1i}/n_{2i})$ 'offset' the model equation; $\log(\psi)$ plays the role of the grand mean.

Since $\hat{\psi}_{ML}$ is constrained to be positive and has a rather skew distribution, it is more appropriate to develop the normal approximation on the log scale. The asymptotic standard error of $\hat{\beta}_{ML} = \log(\hat{\psi}_{ML})$ is

$$\text{S.E.}(\beta_{ML}) = 1/\{\Sigma\, d_i\, \hat{p}_i\, \hat{q}_i,\}^{1/2} \qquad (3)$$

where $\hat{p}_i = 1 - \hat{q}_i = \hat{\psi}_{ML} n_{1i}/(\hat{\psi}_{ML} n_{1i} + n_{2i})$ are the fitted binomial probabilities under H_1. For the data in Table 2 we find $\hat{\psi}_{ML} = 1.4255$, $\hat{\beta}_{ML} = 0.3545$ and S.E. $(\hat{\beta}_{ML}) = 0.1073$. Ninety per cent confidence limits for the common rate ratio are thus $\hat{\psi}_{ML}\exp\{\pm 1.645 \times \text{S.E.}(\hat{\beta}_{ML})\} = (1.195, 1.701)$, which may be contrasted with the test based[10] limits of (1.196, 1.699) found by R&B. Although they give virtually identical results for these data, the test based limits are known to be incorrect in some settings[11,12] and are perhaps best reserved for situations where no valid elementary limits are available.

MANTEL-HAENSZEL ESTIMATE AND VARIANCE

The major disadvantage of the maximum likelihood estimate is that it is only implicitly defined as the solution to an equation. Fortunately, as noted by R&B, the robust Mantel-Haenszel estimate is available as an elementary alternative. This is

$$\hat{\psi}_{MH} = \frac{\Sigma R_i}{\Sigma S_i} = \frac{\Sigma d_{1i}\, n_{2i}/N_i}{\Sigma d_{2i}\, n_{1i}/N_i}, \qquad (4)$$

where R_i and S_i are defined by the numerator and denominator expressions, respectively. For the data in Table 2 we find $\hat{\psi}_{MH} = 1.4247$, which is almost the same as the iterative estimate. In fact, $\hat{\psi}_{MH}$ arises as an approximation to the maximum likelihood estimate that is especially good for rate ratios near unity.[13,14]

A robust variance for the Mantel-Haenszel estimate for cohort studies is easily derived.[15] Writing $\hat{\psi}_{MH} - \psi = \Sigma\,(R_i - \psi S_i)/\,\Sigma S_i$ and noting that $E(R_i) = \psi E(S_i)$ under H_1, the asymptotic variance is $\text{Var}_A(\hat{\psi}_{MH}) = \Sigma E(R_i - \psi S_i)^2/\{\Sigma E(S_i)\}^2$. It follows that

$$\mathrm{S.E.}(\hat{\beta}_{MH}) = \hat{\psi}^{-1}_{MH}\,\mathrm{S.E.}(\hat{\psi}_{MH}) = \frac{\{\Sigma n_{1i} n_{2i} D_i/N_i^2\}^{1/2}}{\{\hat{\psi}_{MH}\}^{1/2}\Sigma \dfrac{n_{1i}\, n_{i2}\, D_i}{N_i(\hat{\psi}_{MH} n_{1i} + n_{2i})}} \qquad (5)$$

This gives $\mathrm{S.E.}(\hat{\beta}_{MH}) = 0.1074$ for the data in Table 2. Formula (5) has an advantage over other variance estimates[13] in that it depends on d_{1i} and d_{2i} only through $\hat{\psi}_{MH}$. However, the analogous formula for sets of 2×2 tables as arise in case-control studies is considerably more complicated. The exact variances $E(R_i - \psi S_i)^2$ are not so easily obtained in that case and therefore have either been approximated[16,17] or replaced with the empirical quantities $(R_i - \hat{\psi}_{MH} S_i)^2$.[18]

Our experience with these and other cohort data is that the Mantel-Haenszel and maximum likelihood estimates are extremely close even when ψ departs from one. This is easy to check, moreover, by substituting $\hat{\psi}_{MH}$ into the estimating equation (2). If the two sides differ by more than a per cent, say, a one-step correction to $\hat{\beta}_{MH}$ is available as

$$\hat{\beta}_C = \hat{\beta}_{MH} + \frac{\Sigma d_{1i} - \Sigma D_i \hat{\psi}_{MH} n_{1i}/(\hat{\psi}_{MH} n_{1i} + n_{2i})}{\Sigma D_i\, \hat{p}_i\, \hat{q}_i} \qquad (6)$$

The correction is unnecessary in the present example since $\Sigma d_{1i} = 630$ and $\Sigma \hat{\psi}_{MH} D_i n_{1i}/(\hat{\psi}_{MH} n_{1i} + n_{2i}) = 629.9487$ are so close. Nevertheless, in order to illustrate its application, we use $\hat{\psi}_{MH}$ to calculate the fitted probabilities and then $\Sigma D_i \hat{p}_i \hat{q}_i = 86.7729$ and find

$$\hat{\beta}_C = 0.35395 + \frac{630 - 629.9487}{88.7729} = 0.35454,$$

which agrees with $\hat{\beta}_{ML}$ to the number of decimal places shown.

In large samples the ratio of (3) to (5) tends to a quantity which is less than one unless $\psi = 1$. Thus there is some loss of efficiency with the Mantel-Haenszel estimate under the alternative hypothesis.[19,20] However, our experience is that the two standard errors are usually close, though not always so close as for the example here. Thus the loss of efficiency appears to be rather slight, as is already known for case-control studies.[15]

TESTING FOR HETEROGENEITY AND TREND IN THE RATE RATIOS

The right hand column of Table 2 indicates a steady decline in the coronary death rate ratios for smokers versus non-smokers with advancing age, and there is substantial question as to whether the data are adequately represented by a single summary ratio. Fitted values $D_i \hat{p}_i$ and $D_i \hat{q}_i$ calculated under the hypothesis of a common ratio deviate markedly from the observed values in the youngest and oldest age groups (Table 2). These deviations may be inserted in the usual chi-square formula

$$\chi^2_{I-1} = \sum \frac{(d_{1i} - D_i \hat{p}_i)^2}{D_i \hat{p}_i \hat{q}_i} =$$

$$\sum \frac{(d_{1i} - D_i \hat{p}_i)^2}{D_i \hat{p}_i} + \frac{(d_{2i} - D_i \hat{q}_i)^2}{D_i \hat{q}_i} \qquad (7)$$

to test H_1 against the alternative of general heterogeneity. For the data in Table 2, we find $\chi^2_4 = 11.15$ on $I - 1 = 4$ degrees of freedom ($p = 0.026$) which may be compared to the likelihood ratio test value 12.13 found by R&B.

When there is a natural ordering of the strata, as for age in this example, a more powerful test of H_1 is given by the following modification of the usual test for a trend in proportions[21] which arises as the score test of H_2 vs H_1:

$$\chi^2_1 = \frac{\{\Sigma\, x_i(d_{1i} - D_i \hat{p}_i)\}^2}{\Sigma x_i^2 D_i \hat{p}_i \hat{q}_i - (\Sigma\, x_i D_i \hat{p}_i \hat{q}_i)^2/\Sigma D_i \hat{p}_i \hat{q}_i} \qquad (8)$$

This statistic, which takes advantage of any systematic change in the deviations $d_{1i} - D_i \hat{p}_i$ with the stratification variable, is referred to tables of chi-square with only one degree of freedom. Setting $x_i = i$ for $i = 1, 2, \ldots, 5$ for use with the Table 2 data, we find $\chi^2_1 = (-34.965)^2/118.7 = 10.30$ ($p = 0.001$) and conclude that most of the heterogeneity in the observed age-specific ratios is due to a linear trend with age. In fact, the goodness-of-fit chi-square for the model H_2 with $f(\theta x) = \exp(\theta x)$, as obtained from a GLIM analysis, is $\chi^2_2 = 1.44$ (NS). R&B show that the data are also consistent with an additive effect of smoking on the age-specific rates.

DISCUSSION

The preceding has demonstrated that simple and efficient statistical methods are available for the comprehensive analysis of incidence density data in cohort studies. The fitted frequencies used in the tests for heterogeneity and trend should be found by maximum likelihood. Fortunately, the maximum likelihood estimate $\hat{\beta}_{ML}$ may be obtained in one or two iterations using (6) and $\hat{\beta}_{MH}$ as a starting value.

Extensions of these basic techniques may be made to accommodate an exposure variable that has several ordered levels. For example, Hakulinen[23] provides the appropriate generalization of (1) for testing the null hypothesis against the alternative of increasing incidence with increasing exposure. Indeed, all the methods presented in Section 4.5 of Breslow and Day[24] for analysis of case-control data in a series of $2 \times K$ tables may be adapted for use with incidence density data in a similar fashion to that shown here.

ACKNOWLEDGEMENT

Research supported in part by grants from the US National Cancer Institute and the Alexander von Humboldt Foundation.

REFERENCES

[1] Doll R and Hill A B. Mortality of British doctors in relation to smoking: observations on coronary thrombosis. *Nat Cancer Inst Monogr* 1966; **19:** 205–68.

[2] Rothman K J and Boice J D. Epidemiologic Analysis with a Programmable Calculator. NIH Publication 79-1649. Washington, US Government Printing Office, 1979.

[3] Cox D R and Hinkley D V. Theoretical Statistics. London, Chapman and Hall, 1974.

[4] Day N E and Byar D. Testing hypothesis in case-control studies—equivalence of Mantel-Haenszel statistic and logit score tests. *Biometrics* 1979; **35:** 623–30.

[5] Cochran W G. Some methods of strengthening the common χ^2 tests. *Biometrics* 1954; **10:** 417–51.

[6] Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Nat Cancer Inst* 1959; **22:** 719–48.

[7] Shore R E, Pasternack B S and Curnen M G. Relating influenza epidemics to childhood leukemia in tumor registries without a defined population base: a critique with suggestions for improved methods. *Am J Epidem* 1976; **103:** 527–35.

[8] Gart J J. The analysis of ratios and cross-product ratios of Poisson variates with applications to incidence rates. *Commun Stat Theory Meth* 1978: **A7:** 917–37.

[9] Baker R J and Nelder J A. The GLIM System. Release 3. Oxford, Numerical Algorithms Group, 1978.

[10] Miettinen O S. Estimability and estimation in case-referrent studies. *Am J Epidem* 1976; **103:** 226–35.

[11] Halperin M. Letter to the Editor. *Am J Epidem* 1977; **105:** 496–8.

[12] Brown C C. The validity of approximate methods for interval estimation of the odds ratio. *Am J Epidem* 1981; **113:** 474–80.

[13] Tarone R E. On summary estimators of relative risk. *J Chron Dis* 1981; **34:** 463–8.

[14] Clayton D G. The analysis of prospective studies of disease aetiology. *Commun Stat Theory Meth* 1982; **11:** 2129–55.

[15] Breslow N. Odds ratio estimators when the data are sparse. *Biometrika* 1981; **68:** 73–84.

[16] Hauck W W. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* 1979; **35:** 817–20.

[17] Ury H. Hauck's approximate large-sample variance of the Mantel-Haenszel estimator. *Biometrics* 1982; **38:** 1094–5.

[18] Breslow N E and Liang K Y. The variance of the Mantel-Haenszel estimator. *Biometrics* 1982; **38:** 943–52.

[19] Nurminen M. Efficient estimators of common relative risk. *Biometrika* 1981; **68:** 525–30.

[20] Tarone R E, Gart J J and Hauck W W. On the asymptotic inefficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika* 1983; **70:** 519–22.

[21] Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11:** 375–86.

[22] Tarone R E and Gart J J. On the robustness of combined tests for trends in proportions. *J Am Stat Ass* 1980; **75:** 110–16.

[23] Hakulinen T. A Mantel-Haenszel statistic for testing the association between a polychotomous exposure and a rare outcome. *Am J Epidem* 1981; **113:** 192–7.

[24] Breslow N E and Day N E. Statistical Methods in Cancer Research I: The Analysis of Case Control Studies. IARC Scientific Publications No. 32. Lyon, IARC, 1980.

(Revised version received September 1983)