Epidemiologic Perspectives & Innovations

BioMed Central

Methodology

Open Access

An easy approach to the Robins-Breslow-Greenland variance estimator

Paul Silcocks*

Address: Trent Research & Development Support Unit, Medical School, Queen's Medical Centre, Nottingham, NG7 2UH UK

Email: Paul Silcocks* - paul.silcocks@nottingham.ac.uk

* Corresponding author

Published: 26 September 2005

Epidemiologic Perspectives & Innovations 2005, **2**:9 doi:10.1186/1742-5573-2-9

Received: 26 April 2005

Accepted: 26 September 2005

This article is available from: http://www.epi-perspectives.com/content/2/1/9

Abstract

The Mantel-Haenszel estimate for the odds ratio (and its logarithm) in stratified case control studies lacked a generally acceptable variance estimate for many years. The Robins-Breslow-Greenland estimate has met this need, but standard textbooks still do not provide an explanation of how it is derived. This article provides an accessible derivation which demonstrates the link between the Robins-Breslow-Greenland estimate and the familiar Woolf estimate for the variance of the log odds ratio, and which could easily be included in Masters level courses in epidemiology. The relationships to the unconditional and conditional maximum likelihood estimates are also reviewed.

Introduction

The Mantel-Haenszel (MH) estimate for the summary odds ratio across several 2 × 2 tables, ψ_{MH}, was proposed in 1959 [1]. Over twenty years later the lack of a robust estimate for its variance was still being noted [2], yet only a few years afterwards, Robins, Breslow and Greenland introduced their now generally-accepted variance estimator [3] for the Mantel-Haenszel log-odds ratio (denoted by the RBG estimate). This replaced estimation of confidence limits based on the unsatisfactory test-based procedure of Miettinen or the computationally intensive Cornfield type limits which had hitherto been used.

While a useful review of Mantel-Haenszel methods has been published, including some aspects of the historical development towards the RBG estimator [4] the formal derivations by Robins, Breslow & Greenland [3] and Phillips & Holland [5] are not, in the view of this author, easily comprehended. The former omits steps in the argument, while the latter appeals to descending factorial powers. Possibly it is no surprise that even modern textbooks [6,7] merely state the RBG formula without deriving it.

While other variance estimators exist, some are ad hoc, such as the application of the cohort study formula to case-control data suggested by Clayton and Hills [8], only apply to the large few strata case [9] or are closely related to the RBG estimator [10]. One rather different exception is Sato's formula [11] but this procedure gives confidence limits directly in the odds ratio scale.

It is the intention of this article to present an informal derivation of the RBG estimator as an extension of the familiar variance formula of Woolf [12], and which could readily be included in standard textbooks of epidemiol-

ogy or biostatistics. I will describe this from the perspective of a case-control study.

Analysis

How does the Mantel-Haenszel estimate arise?

Consider a stratified case-control study for which the i[th] of k independent tables is:

	Case	*Control*
Exposed	a_i	b_i
Unexposed	c_i	d_i
Total	n_{1i}	n_{0i}

Neglecting constants, the unconditional likelihood for the i[th] table is:

$$Lik = \theta_i^{a_i}(1-\theta_i)^{c_i}\phi_i^{b_i}(1-\phi_i)^{d_i}$$

where in the i[th] table θ_i = probability of exposure if a case and $\varnothing_i$ = probability of exposure if a control.

The maximum likelihood estimate (MLE) for $\varnothing_i$ is given by $b_i/(b_i + d_i)$ and if we re-parameterise θ_i as $\psi\varnothing_i/[\psi\varnothing_i + (1 - \varnothing_i)]$, where ψ is the odds ratio (assumed common to all tables), the contribution to the overall log likelihood made by terms involving ψ is:

$$\sum a_i \ln\{\psi\varnothing_i/[\psi\varnothing_i + (1 - \varnothing_i)]\} + c_i \ln\{1/[\psi\varnothing_i + (1 - \varnothing_i)]\}.$$

Differentiating with respect to ψ and equating to zero, and rearranging (noting that $a_i + c_i = n_{1i}$)

we obtain:

$$\sum a_i \ln\{\psi\varnothing_i/[\psi\varnothing_i + (1 - \varnothing_i)]\} + c_i \ln\{1/[\psi\varnothing_i + (1 - \varnothing_i)]\}.$$

i.e., $\sum\{a_i - n_{1i}\psi\varnothing_i/[\psi\varnothing_i + (1 - \varnothing_i)]\} = 0$

i.e., $\sum\{[\psi a_i\varnothing_i + a_i - a_i\varnothing_i - n_{1i}\psi\varnothing_i]/[\psi\varnothing_i + (1 - \varnothing_i)]\} = 0.$

This must be solved numerically to obtain the MLE for ψ, but if the denominators do not vary too much across the tables we merely have to solve:

$$\sum[\psi a_i\varnothing_i + a_i - a_i\varnothing_i - n_{1i}\psi\varnothing_i] = 0$$

i.e., $\sum[\psi(a_i - n_{1i})\varnothing_i + a_i(1 - \varnothing_i)] = 0$

or, $\sum a_i(1 - \varnothing_i) = \sum \psi(n_{1i} - a_i)\varnothing_i$

giving, $\sum a_i(1 - \varnothing_i) = \psi\sum(n_{1i} - a_i)\varnothing_i$

and since, $\psi_i = b_i/(b_i + d_i) = b_i/n_{0i}$

$$\hat{\psi} = \frac{\sum a_i d_i / n_{0i}}{\sum b_i c_i / n_{0i}}.$$

This can be used as a first approximation to find the MLE (if there is only one table then ψ *is* the unconditional MLE = ad/bc). Now in stratified case-control studies with a constant ratio, *r*, of controls to cases, the total number of subjects in each stratum is given by $n_i = n_{0i}(1 + r)$, so $n_{0i} = n_i/(1 + r)$. A constant *r* will be achieved by design if there is caliper matching; otherwise – as with a post-stratified analysis – this will be only approximately true. The term $(1 + r)$ can then be cancelled and we are left with:

$$\hat{\psi} = \frac{\sum a_i d_i / n_i}{\sum b_i c_i / n_i} = \psi_{MH}.$$

The MH estimator is therefore a first approximation to the unconditional MLE in the large strata case with a constant control:case ratio across strata. However the MH estimator actually *coincides* with the *conditional* MLE for the matched pairs design, as outlined, for example on page 164 of Breslow & Day [2].

The sensitivity to variation in the $\varnothing_i$ and constancy of the control:case ratio is not high, as shown by the data in Table 1. In a sense this would be expected because for the most sparse (e.g., pair-matched) data the control:case ratio will be constant, and while the $\varnothing_i$ then have maximum variance – being only 0 and 1, the MH estimate coincides with the conditional MLE. Conversely, for large strata the control:case ratio will vary, but the variance of the $\varnothing_i$ will be less and the MH estimate will then approximate the unconditional MLE.

Deriving the variance of the Mantel-Haenszel estimate

Consider again the i[th] 2 × 2 table, giving the frequencies in each cell:

	Case	*Control*
Exposed	a_i	b_i
Unexposed	c_i	d_i

For odds ratio $\hat{\psi}$, estimated for a single table by the cross-product ratio $a_i d_i/b_i c_i$, application of the delta method gives Woolf's logit-based formula [8]:

Table 1: Simulated case-control data with true odds ratio = 5

Case	Control	ø	Controls:cases
36	97	0.58	4
6	71		
42	168		
Ca	Co		
41	79	0.94	2
1	5		
42	84		
Ca	Co		
2	1	0.02	2
26	55		
28	56		
Ca	Co		
19	25	0.30	3
9	59		
28	84		
Ca	Co		
20	41	0.37	4
8	71		
28	112		
Ca	Co		
30	21	0.62	1
4	13		
34	34		
Ca	Co		
22	26	0.46	2
6	30		
28	56		

Odds ratio estimates (Stata v7.0):
Mantel-Haenszel 4.38 (95% CI 2.85 to 6.72)
Conditional MLE 4.36 (95% CI 2.85 to 6.67)
Unconditional MLE 4.42 (95% CI 2.88 to 6.78)

$\text{var}(\psi) \approx \psi^2(1/a_i + 1/b_i + 1/c_i + 1/d_i)$ with $n_i = a_i + b_i + c_i + d_i$ and, $\text{var}[\ln(\hat{\psi})] \approx (1/a_i + 1/b_i + 1/c_i + 1/d_i)$

The delta method is a widely used procedure in statistics when an approximation is needed for the variance of a function of a variable whose variance is known. In this instance the variable with known variance is a proportion p, and the function is the logit. The basic delta method formula is: $\text{var}(y) \approx (dy/dx)^2 \text{var}(x)$ from which, if $y = logit(p = \ln[p/(1 - p)]$,

$\text{var}(y) \approx (1/p + 1/(1 - p))^2 p(1 - p)/n$

$= (1/p + 1/(1 - p))1/n$

$= (1/a + 1/b)$

if $p = a/n$ and $n = a + b$.

Here we have two independent proportions (the proportion of cases and controls exposed) and Woolf's formula is obtained by estimating the variances of the separate logits and adding them.

For k such 2 × 2 tables, each representing a separate stratum, the Mantel-Haenszel pooled estimate of the *common* odds ratio ψ is given by:

$$\psi_{MH} = \frac{\sum_i a_i d_i / n_i}{\sum_i b_i c_i / n_i}$$

$$= \frac{\sum_i (b_i c_i / n_i)\psi_i}{\sum_i b_i c_i / n_i}$$

Hence ψ_{MH} is a weighted average of the stratum-specific odds ratios. The weights approximate the inverse of the variance of each $\hat{\psi}_i$ if the true value of $\psi = 1$. Note that the assumption here of a common odds ratio is not required for the Mantel-Haenszel *test*.

To derive the variance, in addition to the approximation involved in application of the delta rule, an assumption is also made that each stratum-specific odds ratio is close enough to the Mantel-Haenszel pooled estimate to permit terms like $a_i d_i / b_i c_i$ to be replaced by ψ_{MH}.

We then proceed by obtaining an approximation which avoids zeros in the formula for $\text{var}[\ln(\hat{\psi})]$. The motivation for this can be seen by comparing the weights for ψ_{MH} – which are unaffected by zeros except for deleting such strata – whereas if Woolf's variances were used, the result would be indeterminate if cells with zeros were present.

Taking the weights as constant,

$$\text{var}(\psi_{MH}) = \frac{\sum (b_i c_i / n_i)^2 \text{var}(\hat{\psi}_i)}{(\sum b_i c_i / n_i)^2}$$

$$= \frac{\sum (b_i c_i / n_i)^2 (1/a_i + 1/b_i + 1/c_i + 1/d_i)}{(\sum b_i c_i / n_i)^2}.$$

Assuming a common odds ratio ψ, estimated by ψ_{MH}, this can be written as:

$$\mathrm{var}(\psi_{MH}) \approx \frac{\psi_{MH}^2 \sum (b_i c_i/n_i)^2 (1/a_i + 1/b_i + 1/c_i + 1/d_i)}{(\sum b_i c_i/n_i)^2}.$$

Leading to a formula suggested by Hauck [9]:

$$\mathrm{var}[\ln(\psi_{MH})] \approx \frac{\sum (b_i c_i/n_i)^2 (1/a_i + 1/b_i + 1/c_i + 1/d_i)}{(\sum b_i c_i/n_i)^2}.$$

As mentioned above, a problem with this formula is that it fails if cell entries are zero. However we can proceed further by re-writing the formula as:

$$\mathrm{var}[\ln(\psi_{MH})] \approx \frac{\sum (b_i c_i/n_i)^2 ([a_i + d_i]/a_i d_i + [b_i + c_i]/b_i c_i)}{(\sum b_i c_i/n_i)^2}$$

$$= \frac{\sum (b_i c_i)/n_i^2 \; (b_i c_i)\,([a_i + d_i]/a_i d_i + [b_i + c_i]/b_i c_i)}{(\sum b_i c_i/n_i)^2}.$$

On substituting $1/\psi_{MH}$ for $(b_i c_i/a_i d_i)$:

$$\mathrm{var}_1 \ln(\psi_{MH}) \approx \frac{\sum (b_i c_i)/n_i^2 \; ([a_i + d_i]/\varphi_{MH} + b_i + c_i)}{(\sum b_i c_i/n_i)^2}$$

Now if the rows of the 2 × 2 table are interchanged, the variance stays the same. But a similar argument to that above leads to:

$$\mathrm{var}_2 \ln(\psi_{MH}) \approx \frac{\sum (a_i d_i)/n_i^2 \; ([b_i + c_i]/\psi^*_{MH} + a_i + d_i)}{(\sum a_i d_i/n_i)^2}$$

(Note that the new odds ratio $\psi^*_{MH} = (\sum b_i c_i/n_i)/(\sum a_i d_i/n_i)$ formed by exchanging rows is just $1/\psi_{MH}$.) "The" variance, V, of $\ln(\psi_{MH})$ is therefore taken to be the mean of the two estimates [13] as follows:

Let $R = \sum (a_i d_i/n_i)$ and $S = \sum (b_i c_i/n_i)$. On substituting into the two variance formulae:

$$V = \frac{\sum [R^2 (b_i c_i)/\psi_{MH} n_i^2 + S^2 (a_i d_i)/n_i^2][a_i + d_i + \psi_{MH}(b_i + c_i)]}{2R^2 S^2}.$$

Next, divide the top and bottom by S^2 and move the n_i^2 term outside the brackets to obtain:

$$V = \frac{\sum [\psi_{MH} b_i c_i + a_i d_i][a_i + d_i + \psi_{MH}(b_i + c_i)]/n_i^2}{2R^2}$$

which is eq. 9 in Phillips & Holland [5].

If we now put

$P_i = (a_i + d_i)/n_i$ and $Q_i = (b_i + c_i)/n_i$ with $R_i = a_i d_i/n_i$ and $S_i = b_i c_i/n_i$

then $V = \frac{\sum [\psi_{MH} S_i + R_i][P_i + \psi_{MH} Q_i]}{2R^2}$,

which on multiplying out the brackets, rearranging and noting that $R/S = \psi_{MH}$, gives:

$$V = \frac{\sum R_i P_i}{2R^2} + \frac{\sum (P_i S_i + Q_i R_i)}{2RS} + \frac{\sum S_i Q_i}{2S^2}$$

This is the RBG formula!

When there is only one stratum, this reduces to $(1/a + 1/b + 1/c + 1/d)$ which is the familiar logit based formula of Woolf and which approaches 0 as the sample size increases, assuming a finite true odds ratio. Clearly as the RBG variance estimate is a finite sum of such estimators the RBG estimate will also approach 0, for large strata.

The RBG estimator was derived above on the assumption that the stratum-specific odds ratio estimates could be liberally replaced by the common value, in turn estimated by ψ_{MH}; both assumptions are reasonable with large samples per stratum. However, the success of the RBG formula derives from its being applicable also to the sparse data case.

To see this, consider a matched-pairs case control study. The capital letters denote the frequency of case-control pairs.

		CONTROLS	
		Exposed	*Unexposed*
CASES	*Exposed*	*A*	*B*
	Unexposed	*C*	*D*

In such a study each stratum has only two observations. The table can be decomposed into four types of "unmatched" table according to the exposure category of the case and the control, the frequency of each type being given by the frequency of the corresponding case-control pairs:

	Case	*Control*
Exposed	1	1
Unexposed	0	0

There are A such tables

	Case	*Control*
Exposed	1	0
Unexposed	0	1

There are B such tables

	Case	*Control*
Exposed	0	1
Unexposed	1	0

There are C such tables

	Case	*Control*
Exposed	0	0
Unexposed	1	1

There are D such tables

Only the B such tables with $a_i = d_i = 1$ and the C such tables with $b_i = c_i = 1$ contribute to the estimate of the odds ratio. Note that these are disjoint sets of tables.

Under these circumstances: ψ_{MH} = B/C which coincides with the conditional MLE and:

a) the middle term of the RBG formula vanishes because if $b_i c_i = 1$ then $(a_i + d_i) = 0$, and if $a_i d_i = 1$ then $(b_i + c_i) = 0$

b) $R = \sum a_i d_i / n_i = B/2$ & $S = \sum b_i c_i / n_i = C/2$

c) There are B terms in which $a_i d_i (a_i + d_i) = 2$ C terms in which $b_i c_i (b_i + c_i) = 2$

giving:

$V = B/B^2 + C/C^2 = 1/B + 1/C$

This is not only the familiar logit based formula for the variance of the log odds ratio for matched pairs, but is also the variance of the conditional maximum likelihood estimate. This is asymptotically consistent from the general properties of a MLE (and it's easy to see that as the number of tables increases, $V \to 0$).

In other words, the RBG formula, though derived here without assuming validity in the sparse case, does in fact possess this property.

Table 1 shows how closely the conditional maximum likelihood estimate, unconditional maximum likelihood estimate, and MH estimate agree, despite varying ψ_i and control:case ratio.

Conclusion

The Mantel-Haenszel estimate of the odds ratio approximates the maximum likelihood estimate for large, few strata and coincides with the conditional maximum likelihood estimate for the sparse data (matched pairs) case.

The RBG formula is the estimator of choice for the variance of the Mantel-Haenszel log-odds-ratio because it applies both in the large few strata case and in the many sparse strata case (as in matched pairs analysis), when the RBG variance estimate actually coincides with the conditional maximum likelihood variance estimate.

Moreover the RBG formula reduces to familiar standard forms for a single stratum and for matched pairs.

Formal derivation of the RBG formula is tricky but an informal, accessible derivation is possible as outlined above, which uses nothing more advanced than the delta method for approximating a variance.

Competing interests

The author(s) declare that they have no competing interests.

References

1. Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI* 1959, 22:719-748.
2. Breslow NE, Day NE: *Statistical methods in cancer research, Volume I – the analysis of case-control studies* Lyons: International Agency for Research on Cancer; 1980.
3. Robins J, Breslow N, Greenland S: **Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models.** *Biometrics* 1986, **42:**311-323.
4. Kuritz SJ, Landis JR, Koch GG: **A general overview of Mantel-Haenszel methods: applications and recent developments.** *Ann Rev Public Health* 1988, **9:**123-160.
5. Phillips A, Holland PW: **Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate.** *Biometrics* 1987, **43:**425-431.
6. Armitage P, Berry G, Matthews JNS: *Statistical methods in medical research* 4th edition. Oxford: Blackwell Science; 2002.
7. Fleiss JL, Levin B, Paik MC: *Statistical Methods for Rates & Proportions* Chichester: John Wiley; 2003.
8. Clayton D, Hills M: *Statistical methods in epidemiology* Oxford: Oxford University Press; 1995.
9. Hauck WW: **The large-sample variance of the Mantel-Haenszel estimator of a common odds ratio.** *Biometrics* 1979, **35:**817-819.
10. Flanders WD: **A new variance estimator for the Mantel-Haenszel odds ratio.** *Biometrics* 1985, **41:**637-642.
11. Sato T: **Confidence limits for the Common Odds Ratio Based on the Asymptotic Distribution of the Mantel-Haenszel Estimator.** *Biometrics* 1990, **46:**71-80.

12. Woolf B: **On estimating the relationship between blood group and disease.** *Human Genet* 1955, **19:**251-253.
13. Ury HK: **Hauck's approximate large-sample variance of the Mantel-Haenszel estimator [letter].** *Biometrics* 1982, **38:**1094-1095.