Biometrical Journal **48** (2006) 1, 23–34 DOI: 10.1002/bimj.200510167

An 'Unconditional-like' Structure for the Conditional Estimator of Odds Ratio from 2 × 2 Tables

James A. Hanley[*,1,2] and **Olli S. Miettinen**[1,3,4]

[1] Department of Epidemiology, Biostatistics and Occupational Health, Faculty of Medicine, McGill University, 1020 Pine Avenue West, Montreal, Quebec, H3A 1A2, Canada
[2] Division of Clinical Epidemiology and Biostatistics, Royal Victoria Hospital, Montreal, Quebec, Canada
[3] Department of Medicine, Faculty of Medicine, McGill University, Montreal, Quebec, Canada
[4] Department of Medicine, Weill Medical College, Cornell University, New York, NY, USA

Received 13 February 2005, revised 28 June 2005, accepted 30 July 2005

Summary

In the estimation of the odds ratio (*OR*), the conditional maximum-likelihood estimate (*cMLE*) is preferred to the more readily computed unconditional one (*uMLE*). However, the exact *cMLE* does not have a closed form to help divine it from the *uMLE* or to understand in what circumstances the difference between the two is appreciable. Here, the *cMLE* is shown to have the same 'ratio of cross-products' structure as its unconditional counterpart, but with two of the cell frequencies augmented, so as to shrink the unconditional estimator towards unity. The augmentation involves a factor, similar to the finite population correction, derived from the minimum of the marginal totals.

Key words: Odds ratio; Non-central hypergeometric distribution; Case-control studies.

Abbreviations

OR	odds ratio (parameter)
or	point estimate/estimator of *OR*
	or_u: unconditional *or*
	or_c: conditional *or*
MLE	maximum-likelihood estimate/estimator
	cMLE: conditional *MLE*
	uMLE: unconditional *MLE*
MHE	Mantel-Haenszel estimate/estimator

1 Introduction

Both conditional and unconditional logistic regression are widely used in modern-day biometrical applications. The conditional approach is considerably more complex, both conceptually and computationally, and can sometimes yield an odds ratio estimate that is quite different from that obtained with the unconditional approach. Beginning with the work of Holford, White, and Kelsey (1978) on matched pairs, several authors have shown how software that uses the unconditional approach can be tricked into performing conditional maximum likelihood estimation. Much of the early work (e.g., Tjur (1982), Lindsay, Clogg, and Grego (1991), and Agresti (1993)) focused on Rasch analysis, which is widely used in research on educational testing. In the special case with just one item-difficulty parameter, this parameter corresponds to the log odds-ratio parameter that is the focus in biomedical

[*] Corresponding author: e-mail: James.Hanley@McGill.CA, Phone: +1 514 398 6270, Fax: +1 514 398 4503

applications of conditional logistic regression. More recently, authors such as Neuhaus, Kalbfleisch, and Hauck (1994), Lindsey (2000), and Rice (2004) have focused on the matched case-control study, using the now-better-understood deep connections between Rasch, latent class and mixture models and the estimates obtained from the conditional approach.

These various insights are based, for the most part, on clever representations in regression models, and on mathematically sophisticated arguments. Thus they are less accessible to epidemiologists, especially those who like to visualize their data as a series of 2 × 2 tables and – when possible – use the classic Mantel–Haenszel summary odds ratio estimator for stratified data. We begin with a single 2 × 2 table since it serves as a more useful way to illustrate the differences, and similarities, between the conditional and unconditional maximum likelihood estimators of odds ratio; the stratified counterpart is a natural extension of this.

The analysis of a single 2 × 2 table is usually represented as the comparison of two binomials with unknown parameter values P_1 and P_0, often with focus on the odds ratio, $OR = \frac{P_1}{1 - P_1} \Big/ \frac{P_0}{1 - P_0}$, as the comparative parameter of interest. More generally, with J pairs of binomials with parameters P_{1j} and P_{0j}, $j = 1, \ldots, J$, interest may focus on the presumed common OR across the pairs or strata, on $OR_j \equiv OR$.

In the context of a single pair of binomials, the data are laid out as a single 2 × 2 table; and it has become commonplace to denote the entries in the first row as a and b and those in the second as c and d, a and c forming the first column. The rows and columns are so defined that the point estimate p_1 of P_1 involves a as its numerator input.

With this notation, the familiar point estimator of OR is the cross-products ratio ad/bc; and the familiar asymptotic standard error (*SE*) of the logarithm of this is $\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{1/2}$. These statistics are derived by considering two *independent binomials*, say the distributions of the frequencies in the "*a*" and "*c*" cells conditional on the respective row (first, or "fixed," margin) totals but *unconditional* on the column (second, or "free," margin) totals. That point estimator is the "unconditional" maximum-likelihood estimator, *uMLE*.

Given the fourth 2 × 2 table in Table 1, many researchers would readily estimate the odds ratio as $\frac{3 \times 3}{1 \times 1} = 9.0$, without appreciating that there is an important alternative to this unconditional approach. This is to *condition* on the second margin as well. Now the interest in two binomial realizations gets to be replaced by focus on a single cell, usually the frequency in the "*a*" cell, as the realization for an *extended-hypergeometric* variate. Its distribution, conditional on all of the margins, is determined by the *OR* parameter alone.

The corresponding exact "conditional" *MLE*, the *cMLE*, of *OR* has no general closed form. Therefore the estimate generally has to be found as an iterative solution of an polynomial equation. This lack of a closed form is unfortunate for reasons well beyond calculational demands, since the *cMLE* is, on theoretical grounds, the preferable estimate. With the intra-stratum information sparse, the *uMLE* tends to be too extreme: in the limiting situation of matched pairs, where the *cMLE* is the familiar ratio of the numbers of discordant pairs, the *uMLE* is the square of this (Breslow and Day, 1980, p. 250).

Faced with the stratified data in Table 1, epidemiologists who wish to be "close to their data" would readily calculate the stratum-specific *uMLE*'s but would be unable to divine from them the values of the corresponding stratum-specific *cMLE*'s. Few statisticians, let alone epidemiologists, could anticipate that in the example cited above the conditional counterpart of the unconditional 9.0 is as low as 6.4. The numerical investigations by Breslow and Day (pp. 250–252) focus on the behavior of the *uMLE*, and so they provide less guidance on how, in specific configurations, the *cMLE* can be 'triangulated' from the easily calculated *uMLE*. Thus, having a suitable closed form for the *cMLE* would help in preliminary data analyses; it would also help explain on a theoretical level the data circumstances where the difference between the two estimates is appreciable.

Table 1 Accuracy of (stratum-specific and summary) conditional point estimates *or* of *OR* based on null f (expression (1)) and refined f (Appendix 2), and of the corresponding *SE*s of log(*or*), for six 2×2 tables representing six strata.

Table[1]	f_0	or_{f_0}	or_{f_r}	or_{exact}	*uMLE*	SE_{f_0}	SE_{f_r}	SE_{exact}[2]	$SE(uMLE)$[3]
2 *1* 1 *3*	$\frac{2}{9}$	4.46	4.45	4.45	6.00	1.51	1.49	1.49	1.68
1 *1* 1 *6*	$\frac{7}{16}$	4.48	4.58	4.58	6.00	1.57	1.59	1.59	1.78
12 *1* 2 ***1***	$\frac{7}{15}$	5.05	5.10	5.10	6.00	1.50	1.51	1.51	1.61
3 *1* 1 *3*	$\frac{1}{7}$	6.60	6.41	6.41	9.00	1.49	1.45	1.44	1.63
4 *1* 1 *4*	$\frac{1}{9}$	11.4	10.9	10.9	16.0	1.44	1.39	1.39	1.58
7 *3* 3 *7*	$\frac{1}{19}$	4.96	4.95	4.95	5.44	0.94	0.94	0.94	0.98
Summary		5.74	5.75	5.72	7.12[4]	0.54	0.54	0.54	0.59[4]

[1] Cell frequencies are shown with a' **in bold** and b' *italicized.*

[2] Square root of the inverse of the exact extended-hypergeometric variance corresponding to or_{exact}, that is, to the exact *cMLE*.

[3] $(1/a + 1/b + 1/c + 1/d)^{1/2}$ in each stratum

[4] Via unconditional logistic regression with 5 indicator variates for 6 strata.

Towards this end, we here present an instructive representation of the *cMLE* of *OR*, one that still has the same 'ratio of cross-products' structure as its familiar unconditional counterpart.

2 Unstratified Data

Although in practice the conditional approach to *OR* estimation is more relevant in the context of combining information from *several* 2×2 tables, such as the six illustrative ones in Table 1, or the 12 and the 58 tables from the two actual studies in Table 2, we begin with the case of a *single* table, where the proposed structure is more easily seen.

2.1 Focus/notation, in relation to a *specific* marginal frequency

The usual approach is to focus on cell entries a, b, c, and d and on the corresponding expected frequencies A, B, C, and D. A more effective approach to the conditional assessment of *OR* begins with focus on the (or a) row or column such that the sum of the two cell entries represents the *minimum* of the marginal totals. We denote the cell entries in this by a' and b' and the corresponding marginal total $(a' + b')$ by m (minimum); and in particular, we let a' represent either a or d (and not b or c). We denote by c' and d' the realizations in the other row or column, with d' diagonal to a' and, hence, c' diagonal to b'. Three examples of this notation are shown below. We denote the expected frequency corresponding to a' by A', etc.

$3(d')$	$(c')\ 2$	5
$1(b')$	$(a')\ 1$	$2(m)$
4	3	$7(t)$

$3(d')$	$(b')\ 1$	4
$2(c')$	$(a')\ 1$	3
5	$(m)\ 2$	$7(t)$

$1(a')$	$(b')\ 1$	$2(m)$
$2(c')$	$(d')\ 3$	5
3	4	$7(t)$

2.2 An 'unconditional-like' structure for the cMLE

The theoretical basis for the proposed form for the conditional point estimator of OR derives from a relationship first noted by Birch (1964, Eq. (4)) but usually credited to Mantel and Hankey (1975). The conditional expectations A, B, C, and D and the extended-hypergeometric variance V (of any one of the cell frequencies) are linked to OR via the exact parameter relation

$$OR = \frac{AD + V}{BC + V} = \frac{A'D' + V}{B'C' + V} .$$

We focus on A' etc., rather than A, etc., and re-write the relation as

$$OR = \frac{A'(D' + fB')}{B'(C' + fA')} , \tag{1}$$

where $f = V/A'B'$. Then, substitution of the MLEs $\hat{A}' = a'$, ..., $\hat{D}' = d'$ leads to the general form of the point estimator

$$or = \frac{a'(d' + fb')}{b'(c' + fa')} , \tag{2g}$$

The factor, or fraction, f $(= V/A'B')$ is a function of OR – albeit, as shown in Figure 1, only a weak function, given the choice of a' and b'. A first approximation to it is that implied by the null values $V_0 = \frac{(A' + B')\,(C' + D')\,(A' + C')\,(B' + D')}{t^2(t - 1)}$, $A'_0 = \frac{(A' + B')\,(A' + C')}{t}$ and $B'_0 = \frac{(A' + B')\,(B' + D')}{t}$. Thus, as we will demonstrate below, an approximation – accurate for all practical purposes – to the $cMLE$ of OR generally is

$$or = \frac{a'd'_+}{b'c'_+} ; \tag{2n}$$

$$c'_+ = c' + f_0 a', \qquad d'_+ = d' + f_0 b', \qquad f_0 = \frac{t - m}{m(t - 1)} .$$

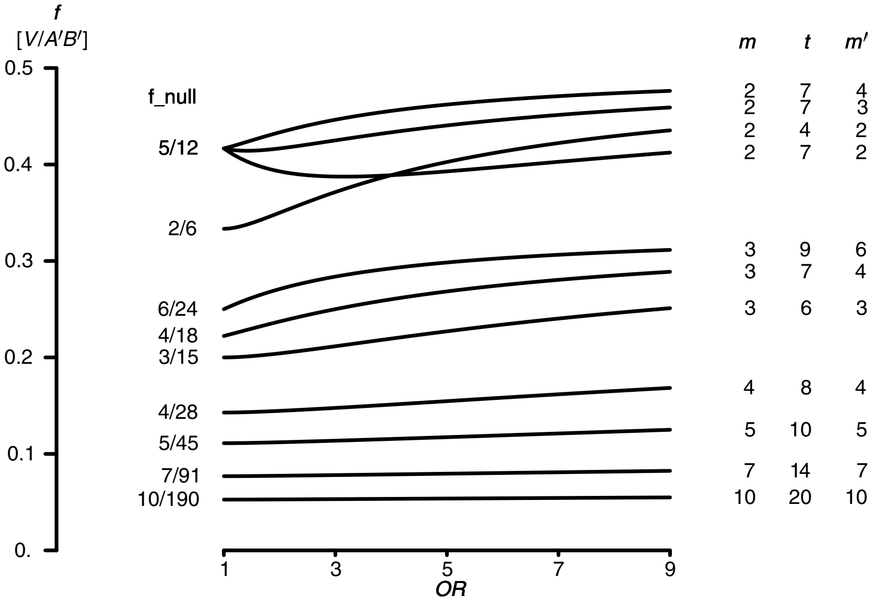

Figure 1 The fraction $f = V/A'B'$ as a function of Odds Ratio OR, for various configurations of m = minimum marginal frequency, t = overall frequency, and $m' = a' + c'$, diagonal to m (see text).

Heuristic justification of this proposition is, first and foremost, a matter of examining whether this estimator has certain familiar properties of the *cMLE*. For a start, when there is no association in the data $(a/b = c/d)$, the exact *cMLE*, just as the *uMLE*, equals unity; and it is immediately obvious that the proposed estimator satisfies this elementary requirement, exactly. Asymptotically (with m infinitely large) the exact *cMLE* coincides with the *uMLE* = ad/bc, and this is satisfied because the asymptotic value of f_0 is zero. And at the other extreme, $m = 1$, where $f_0 = 1$, the proposition implies that, in terms of the expected frequencies A' etc., $OR = \frac{A'(D' + B')}{B'(C' + A')}$, a relation that indeed is familiar from the theory of individually matched data (Miettinen, 1970). Finally, it has been surmised, from examples, and from numerical investigations such as those of Hauck (1984), that the value of the exact *cMLE* is always closer to the null value of $OR = 1$ than the *uMLE* is. Clearly, this relation is inherent in the proposed structure for the *cMLE*.

We illustrate the stratum-specific calculations using the 2 × 2 table employed above to introduce the notation. It is the 'reciprocal' of the table used by Breslow and Day (1980, pp. 125–127) to introduce conditional ML estimation and to illustrate the calculation of the exact *cMLE*. The unconditional estimator yields the simple cross-product ratio

$$uMLE = \frac{3 \times 1}{2 \times 1} = 1.50$$

while the exact *cMLE* is the solution of the polynomial (in this example, quadratic) estimating equation

$$\frac{0 \times \binom{2}{0}\binom{5}{3} OR^0 + 1 \times \binom{2}{1}\binom{5}{2} OR^1 + 2 \times \binom{2}{2}\binom{5}{1} OR^2}{\binom{2}{0}\binom{5}{3} OR^0 + \binom{2}{1}\binom{5}{2} OR^1 + \binom{2}{2}\binom{5}{1} OR^2} = 1\,,$$

i.e., of

$$4 \times OR + 2 \times OR^2 = 2 + 4 \times OR + OR^2\,.$$

The solution is

$$cMLE_{\text{exact}} = 2^{1/2} = 1.41\,.$$

In relation to the minimum marginal frequency, $m = 2$ for this table, the relevant frequencies for the proposed estimator are $a' = 1, b' = 1, c' = 2, d' = 3$ and $f_0 = \frac{5}{12}$, so that the augmented frequencies are $c'_+ = 2 + \left(\frac{5}{12}\right) \times 1$ and $d'_+ = 3 + \left(\frac{5}{12}\right) \times 1$. The proposed crossproduct-like estimator therefore yields

$$cMLE_{\text{approx}} = \frac{1 \times \left(3 + \frac{5}{12}\right)}{1 \times \left(2 + \frac{5}{12}\right)} = 1.41\,.$$

While the exact *cMLE* was easily obtained in this particular instance, in general the estimating equation involves a polynomial of degree m. For example, the estimating equation that yields the *cMLE* of 6.41 involves a 4-th degree polynomial in *OR*. In contrast, the proposed approximation remains a simple cross-product, no matter how large m is.

2.3 cMLE as a 'shrunken' uMLE

The proportional shrinkage in the *cMLE* relative to the *uMLE*, or or_c relative to or_u, is meaningfully expressed by

$$\frac{(or_u - or_c)}{(or_u - 1)}\,,$$

so long as $or > 1$. This measure, in the framework of the formulation of or_c in expression (2g), reduces to the simple form of

$$\frac{fa'}{c' + fa'},$$

so that

$$or_c - 1 = \frac{c'}{c' + fa'} (or_u - 1) . \tag{3}$$

Since f is a weak function of OR (see Figure 1), substitution of the readily computed null value f_0 yields an acceptably accurate measure of the shrinkage. For example, in the Breslow and Day example used to illustrate conditional estimation, the shrinkage factor in the case of the exact *cMLE* is

$$\frac{\sqrt{2} - 1}{1.5 - 1} = 0.8284 ,$$

while that based on f_0 is

$$\frac{c'}{c' + f_0 a'} = \frac{2}{2 + \frac{5}{12} \times 1} = \frac{24}{29} = 0.8276 .$$

Even in the extreme example above, where the *uMLE* is 9.0 and the *cMLE* is 6.4, so that the shrinkage is $5.41/8.0, f_0$ implies $5.60/8.0$.

When $or < 1$, the shrinkage factor is based on the inverses of the estimates. It, in turn, reduces to

$$\frac{fb'}{(d' + fb')},$$

so that

$$\frac{1}{or_c} - 1 = \frac{d'}{d' + fb'} \left(\frac{1}{or_u} - 1\right) . \tag{4}$$

The patterns can be summarized by noting that the discrepancy between the *uMLE* and the *cMLE* is greatest when t is smallest, $a' + c' = t/2$, and the association in the data is substantial.

2.4 Performance of approximations to cMLE

Many users have access to high-level computer packages, and thus to exact *cMLE*'s. Thus, we do not advocate that the computer algorithms that produce exact estimates be replaced with ones that produce approximate ones; rather, our aim is didactic, to give a structure to the conditional estimator, so that the difference between the conditional and unconditional estimates can be appreciated.

It nevertheless is of interest to examine the actual performance of the approximations to the *cMLE*. In Table 1 the proposed point estimator is applied to six separate 2 × 2 tables. The fourth one of these has already been alluded to, while the last one was used by Satten and Kupper (1990) to illustrate a numerically stable algorithm to calculate the exact expected value of the extended-hypergeometric distribution.

The examples represent considerable deviations of the exact *cMLE*s from their corresponding *uMLE*s. Nevertheless, the results produced by the simple point estimator with f_0 as an approximation to the exactly correct value of f, agree quite adequately with the exact *cMLE*s.

The accuracy of the simple approximation using f_0 is at its worst when the discrepancy between the *uMLE* and the *cMLE* is the greatest. Such configurations (see previous section) are represented by the last three examples in Table 1. As illustrated by these examples, if greater accuracy is desired in such situations, a refined approximation (provided in Appendix 2) is possible. The refined estimates are accurate to all three digits of interest in all of the examples.

2.5 Variance of ln of cMLE

Using the same four frequencies a', b', c'_+, d'_+ as are used to obtain the point estimate of *OR*, it can also be shown (see Appendix 1) that an adequately accurate approximation to the *SE* (asymptotic) of the logarithm of the *cMLE* generally is

$$SE_{\log(or)} = [1/a' + 1/b' + (1 - f_0)(1/c'_+ + 1/d'_+)]^{1/2}\,, \tag{5}$$

representing the square root of the inverse of an approximation to the extended-hypergeometric variance evaluated at $A' = a'$. This has a structure similar to that of the *SE* of the logarithm of the unconditional estimate, but with the variance contributions from the two augmented frequencies reduced by a factor based on f_0.

As for the properties of this, we note first that when there is no association in the data $(or = 1)$, this estimator reduces to the square root of the inverse of the (null) hypergeometric variance, exactly. Asymptotically (with m very large), where the distinction between the *cMLE* and the *uMLE* vanishes, this formulation reduces, as it should, to the familiar one involving simply the sum of the inverses of the observed cell-specific frequencies. And when $m = 1$, so that a' and b' are Bernoulli realizations, this formulation reflects the correct variance $A'(1 - A')$. It might be noted also that, notably when m is small, this *SE* is, as it should be, smaller than the familiar unconditional one.

For each *individual* 2×2 table shown in Table 1, the Gaussian approximation to the distribution of a', or equivalently a, does not apply at the limits of *OR*, or even at the point estimate. Thus, even the exact asymptotic *SE* is not helpful in these individual instances. However, an *SE*-based interval may be appropriate if the *OR* is estimated by combining information from several strata (tables), a topic that is considered in the next section.

3 Extension to Stratified Data

In the context of several 2×2 tables, the focus has been on the a's (e.g., Platt, Leroux, and Breslow, 1999). For the value of *OR*, presumed to be common across the strata, there is the corresponding

Table 2 Cell frequencies, redefined, and overall conditional estimates or[1] of *OR* for each of two datasets with individual matching[2].

Dataset	$a' + c'$	a'	Number of such strata	or_{f_0}	or_{exact}	or_{McC}
(i)	1	0	1			
12 strata	1	1	3			
with $m_j = 1$,	2	1	5	22.6	22.6	26.4
and $t_j = 5$	3	1	3			
(ii)	1	0	4			
58 strata	1	1	3			
with $m_j = 1$,	2	0	1			
and $t_j = 5$	2	1	17			
	3	0	1	7.95	7.95	8.11
	3	1	16			
	4	0	1			
	4	1	15			

[1] or_{f_0}: estimate obtained from the proposed approach using f_0.
or_{McC}: McCullagh approximation to *cMLE*.

[2] Sources: (i) Miettinen 1984, p. 151; (ii) Breslow and Day 1980, Appendix III.

expected value, A_j, in stratum j for the conditional distribution (extended-hypergeometric) of the variate whose realization is a_j. By the same token, any given estimate *or* of *OR* implies a corresponding estimate $\hat{A}_j$ of the A'_j. The *cMLE* is (Birch, 1964) that value of *OR* which implies a set of $\hat{A}_j$ values such that

$$\sum \hat{A}_j = \sum a_j \, .$$

Thus, if the six 2 × 2 tables in Table 1 refer to six strata, then at the *MLE* of *OR*, the six implied $\hat{A}_j$ values sum to 29.

To find the root of this estimating equation, one might use the Mantel-Haenszel (1959) estimate (*MHE*) of *OR* as the first trial *or*. Using this *or* value one (i) evaluates the ratios of two polynomials in *or* to calculate the implied values $\{\hat{A}_j\}$, (ii) compares $\sum \hat{A}_j$ with $\sum a_j$, and (iii) revises *or* accordingly before repeating these three steps. Given the numerical difficulties in evaluating these implied expectations exactly, other—necessarily approximate—approaches are sometimes used. For example, for a given estimate *or*, McCullagh (1984) suggested that one obtain the implied value of $\hat{A}_j$ by solving the quadratic equation

$$\frac{\hat{A}_j \hat{D}_j + v_j}{\hat{B}_j \hat{C}_j + v_j} = or \, ,$$

where $v_j = \frac{t_j}{t_j - 1} \, (1/a_j + 1/b_j + 1/c_j + 1/d_j)^{-1}$; $t_j = a_j + b_j + c_j + d_j$; and $\{B_j, \, C_j, \, D_j\}$ are expressed in terms of A_j and the four fixed marginal totals of table *j*. This formulation gives very satisfactory results when each of the four marginal totals in the table exceeds 5 (McCullagh, 1983). The approximation is not, however, very accurate in smaller-margins or extreme-*OR* situations (Breslow and Cologne, 1986).

As we show in the next section, in these situations the approximation can be made more accurate by focusing not on each a_j, but on each a'_j (and its corresponding expectation, A'_j).

3.1 Approximations to cMLE

For any given value of *OR*, there is the corresponding expected value A'_j for the conditional distribution (extended-hypergeometric) of the variate whose realization is a'_j. The *cMLE* is that value, *or*, of *OR* which implies a set of $\hat{A}'_j$ values such that

$$\sum \hat{A}'_j = \sum a'_j \, . \tag{6}$$

Thus, if the six 2 × 2 tables in Table 1 refer to six strata, then at the *cMLE* the implied $\hat{A}'_j$ values sum to 18. Incidentally, it is the location of this observed sum in its range (0 to $\sum m_j$) that determines how appropriate a Gaussian-based confidence interval is.

For any given trial value *or* of *OR*, the corresponding value of $\hat{A}'_j$ is the solution of

$$\frac{\hat{A}'_j \hat{D}'_{+j}}{\hat{B}'_j \hat{C}'_{+j}} = or \, . \tag{7}$$

Upon expressing $\hat{B}'_j$, $\hat{C}'_{+j}$, and $\hat{D}'_{+j}$ in terms of $\hat{A}'$, the marginal totals, and $f_{0j} = \frac{t_j - m_j}{m_j(t_j - 1)}$, this quadratic equation can be solved for $\hat{A}'_j$, separately for each stratum. The solution is

$$\hat{A}'_j = \left| |Q_j| - \left[\frac{Q_j^2 - m_j(a'_j + c'_j) \, or}{(1 - f_{0j}) \, (or - 1)} \right]^{1/2} \right| ;$$

$$Q_j = \frac{[m_j + (a'_j + c'_j)/(1 - f_{0j}) + t_j/(1 - f_{0j}) \, (or - 1)]}{2} \, . \tag{8}$$

With $f_{0j} = 1$ (exactly correct whenever $m_j = 1$) or $or = 1$, the underlying relation is no longer quadratic, and the result is (cf. Breslow and Cologne, 1986)

$$\hat{A}'_j = \frac{(a'_j + c'_j)\ OR}{(a'_j + c'_j)\ OR + b'_j + d'_j}\,,$$

while $or = 1$ implies $\hat{A}'_j = \dfrac{m_j(a'_j + c'_j)}{t_j}$. These particulars are of note because when $f_j = 1$ and/or $or = 1$, the value of expression (8) is indeterminate. The $\hat{A}'_j$ values are summed and compared with $\sum a'_j$, and the trial value of OR is adjusted accordingly. The corresponding new values of $\hat{A}'_j$ are computed; the comparison is repeated, etc., until the trial value of OR produces the equality in expression (6).

3.2 Performance of proposed approximation to *cMLE*

Over the six strata in Table 1, the exact *cMLE* of the common *OR* is 5.72 and its corresponding *SE* of log (*or*) is 0.54. By focusing on the a'_j rather than the a_j, the proposed procedure yields $or = 5.74$ with *SE* of log (*or*) = 0.54. The McCullagh approach yields $or = 5.81$ with *SE* of log (*or*) = 0.56. The *MHE* is 7.07 and its associated *SE* of log (*or*) (Robins, Breslow, and Greenland, 1986) equals 0.59.

Table 2 refers to two datasets resulting from individual matching. The first one, previously cited (Miettinen, 1984, p. 151), concerns the role of induced abortion in the etiology of ectopic pregnancy. Shown are stratum-specific data on matched quintuples (4-for-1 matching). While $m_j = 1$ and $t_j = 5$ in each stratum, informativeness was nevertheless quite variable among the 12 informative strata $(1 \le a' + c' \le 3)$. In 11 of the 12 strata, $a'_j = 1$. The single instance of $a'_j = 0$ was in a stratum with $a' + c' = 1$. The exact *cMLE* is 22.6. For these data, the proposed approach replicates the exact *cMLE* (= 22.6) exactly, with f_0 already. This was to be expected, since the null formulation for f, while generally an approximation, is exact $(f_j = 1)$ irrespective of the value of *OR* in the case of $m_j = 1$. With careful attention to step sizes in the Newton–Raphson process, McCullagh's series of iterative approximations converges, yielding 26.4. Of further interest in this example is the much higher $uMLE = 78.8$, illustrating the danger of fitting a total of 13 parameters to such sparse data. Also of note is the volatility of the result from the *MHE*, namely $or = 33.0$. Had the single instance of $a'_j = 0$ fallen in one of the other two types of stratum, the *MHE* would have been 17.0 or 11.7. The *cMLE* is, of course, independent of the split of the $\sum a'_j$ among the strata.

In the second dataset in Table 2, arising from a study of the effect of exogenous estrogens on the risk of endometrial cancer (Breslow and Day, 1980, p. 162), again, $m_j \equiv 1$ and $t_j \equiv 5$. The proposed approach converges, as expected, to the exact *cMLE* of 7.95, while the McCullagh approximation converges to 8.11, and the *MHE* is 8.46.

3.3 Variance of ln of cMLE

If $0 \ll \sum A'_j \ll \sum m_j$, then the distribution of $\sum a'_j$ is close to Gaussian, and thus one can calculate a reasonably accurate confidence interval based on the *SE* (asymptotic) of the logarithm of the *cMLE*. This *SE*, in turn, can be calculated as $\left(\sum I_j\right)^{-1/2}$, where $I_j = \hat{A}'_j \hat{B}'_j (1 - f_{0j})$.

4 Discussion

4.1 Shrinkage

While it is generally agreed that the *cMLE* is to be preferred to the *uMLE*, it involves extensive computations, even in the case of a single table. Those who wish to inspect stratum-specific *cMLE*s cannot readily calculate them with simple software, or appreciate from the marginal totals how discrepant the *cMLE* and the *uMLE* are, or what are the factors that drive them apart. Nor is it obvious why the conditional estimate is always closer to the null than the unconditional one.

As is evident in Eq. (2*g*), the shrinkage is brought about by augmenting the cell frequencies c' and d'. The augmentations involve the factor f reflecting the size of m, the minimum marginal total – a feature that has not been exploited up to now. The null value of f is directly related to the 'finite population correction factor' in sampling, and to the ratio of the variances of the central-hypergeometric and binomial distributions. We derived its general form from the exact parameter relation linking the conditional expectations A', B', C', and D' and the extended-hypergeometric variance V. This relation's re-expression (expression 1 in Appendix 1) in terms of f clarifies why and by how much the *cMLE* is pulled away from the *uMLE* towards unity, and it shows the critical role of the minimum marginal total, m, in this shrinkage.

While the disparity between the *cMLE* and *uMLE* of *OR* has been investigated numerically (Pike et al., 1980, pp. 250–252), it has not been addressed theoretically. What, then, does expression (3) imply for the case of $or > 1$? For one, that any given value of or_u is associated with maximum disparity when a'/c' is maximal, which occurs when the value of b'/d' is also maximal (as $a'/c' = or_u \times b'/d'$). This situation arises when m/t is maximal, that is, when $m = t/2$. The associated added requirement is, of course, that f be maximal; and in terms of the f_0 in expression (2*n*), it is clear that this occurs when $m = t/2$ is associated with minimal t consonant with the or_u value at issue. With $or > 1$, the proportional degree of augmentation in c'_+ is the key, that in d'_+ when $or < 1$. These conclusions are illustrated by the results in Table 2. The implications of expression (4) for the case of $or < 1$ follow by analogy, with or_u replaced by its inverse.

4.2 Why and when are the approximations accurate?

Many widely available statistical packages can calculate exact *cMLE*'s for datasets much more extensive that ours, and can do so to more decimal places than are needed in practice, all in a fraction of a second. Although our acceptably accurate approximations can be computed even with a hand calculator or spreadsheet, accuracy *per se* was not our aim. Rather, our purpose was to clarify *why* and by *how much* the *cMLE* is pulled away from the *uMLE* and towards unity. Nevertheless, the accuracy of the approximations also is of some theoretical interest.

How does the accuracy of McCullagh's approximation, and of the one proposed here, come about? Both approaches are founded on the parametric relation of

$$OR = \frac{AD + V}{BC + V} .$$

In the context of a single table, McCullagh suggests replacing the expected values by the respective realizations (a, b, c, d); and for the variance estimate he suggests using the asymptotic one multiplied by $\frac{t}{(t-1)}$. This variance estimate is the source of the approximation in his estimator.

We replace the challenge of estimating the extended-hypergeometric variance by that of estimating $f = \frac{V}{A'B'}$ (Appendix 1). On the surface, this appears to adduce nothing new: we, too, appear to face the estimation of V, followed simply by the division of this estimate by the $a'b'$ product; and were one to do this and use McCullagh's estimator of V, our results would be the same as his. The novelty arises from our exploitation of the fact that we have *four possible choices for the definition of* f: $\frac{V}{AB}, \frac{V}{CD}, \frac{V}{AC}$ and $\frac{V}{BD}$. When, in this framework, we define f in terms of that pair (A', B') of cell expectations whose sum represents the minimum, m, of marginal totals, we bypass the challenge of estimating V: we invoke the premise, justified by the patterns in Figure 1, that V is approximately proportional to $A'B'$, so that

$$\frac{V}{A'B'} \approx \frac{V_0}{A'_0 B'_0} = \frac{t - m}{m(t-1)} = f_0 .$$

This is fully accurate whenever $m = 1$, while the McCullagh estimate of V is not; and the greater accuracy extends to $m > 1$ also. The McCullagh formulation, based on *ad* and *bc* augmented by v, could be improved by using for v the formulation involved in expression (5).

As an illustration of the importance of the focus an a' and b' adding up to m, consider the case of $(a, b, c, d) = (6, 1, 1, 1)$. With (a', b', c', d') as $(1, 1, 1, 6)$, expression ($2n$) yields $or = 4.48$, while the exact *cMLE* is 4.58 (Table 1). If, however, expression ($2n$) were used with a in place of a', $a + b$ in place of m, and $c + d$ in place of $t - m$, the result would be 5.12. The McCullagh estimator involves $v = 0.36$, and the result is $or = 4.69$. The formulation in expression (5) implies $v = 0.42$, corresponding to $or = 4.53$, closer to the exact *cMLE* of 4.58.

Acknowledgements This work was supported by operating and equipment grants to JAH from the Natural Sciences and Engineering Research Council of Canada.

5 Appendix 1: Non-null (Extended) Hypergeometric Variance

We wish to express the extended-hypergeometric variance V (of any one of the cell frequencies) in terms of f and the conditional expectations A', B', C', and D'. We begin from the exact relation

$$OR = \frac{A'(D' + fB')}{B'(C' + fA')}\,, \tag{1}$$

where $f = \dfrac{V}{A'B'}$.

The non-null V can then be derived from the extended-hypergeometric probability of a', with the *OR* in it given the structure of that in expression (1). The inverse of V— information on the parameter A' — is the negative of the expectation of the second derivative, with respect to A', of the logarithm of this probability. Writing the probability first as a function of *OR* and then applying the chain rule allows V^{-1} to be expressed as OR'/OR, where OR' denotes the derivative of *OR* with respect to A'. Expressing *OR* in terms of the marginal totals, applying the chain rule several times, and recognizing that f itself is a function of A', leads after some tedious algebra to the expression

$$V^{-1} = \frac{1}{A'} + \frac{1}{B'} + \frac{1-f}{C'_+} + \frac{1-f}{D'_+} - \frac{f't(A' - A'_0)}{C'_+ D'_+}\,, \tag{2}$$

where $f' = \dfrac{\mathrm{d}f}{\mathrm{d}A'}$.

6 Appendix 2: Refinements

As was set forth in Appendix 1, the propositions in expressions ($2n$) and (5) in the text involve the null value f_0 of $f = \dfrac{V}{A'(m - A')}$, that is, the hypergeometric (null) variance together with A'_0 in place of A'. Those simple formulations do not involve notable inaccuracy on this basis because, due to the definitions of a' and b', V remains quite closely proportional to $A'(m - A')$ as A' moves away from A'_0. Another source of some inaccuracy is the f' term, proportionately quite small, that is involved in the exact formulation of V (expression (2), Appendix 1) but is omitted in expression (5).

The "refined" results in Tables 1 and 2 are based on taking the extended-hypergeometric variance to be, as a closer approximation,

$$V = \frac{1}{[1/A' + 1/B' + (1-f)\,(1/C'_+ + 1/D'_+)\,(1-F)]}\,;$$

$$F = \frac{f^2(2A' - m)\,(A' - A'_0)\,t/C'_+D'_+}{1 - f^2A'B'(1/C'_+ + 1/D'_+)}\,.$$

Use of f_0 for the f in this, together with a' for A' etc., yields a first approximation $\hat{V}_1$ for $\hat{V}$. Then, $f_1 = \dfrac{\hat{V}_1}{a'b'}$ leads to $\hat{V}_2$, etc., until convergence. The final approximation for f is used as a replacement for f_0 in expression $(2n)$ for the point estimate of OR; and the SE of $\log(or)$ becomes the square root of the inverse of $fa'b'$.

The variance approximation above involves an approximation to a very complex f' in the exact variance (expression (2), Appendix 1). For its derivation, differentiated is, of course, $f = \dfrac{V}{A'(m - A')}$, but the f' term in V (exact) is omitted. Moreover, in the differentiation, $1/C'_+ + 1/D'_+$ is treated as a constant, different from $A'(m - A')$ and $1 - f$.

These refinements of the simple propositions in Sections 2.1 and 2.4 gain some potential relevance in instances characterized by $m > 1$ (so that $f < 1$) yet m that is small (being far from asymptotic case) and m/t that is large (so that the binomial approximation is inapplicable).

References

Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scandinavian Journal of Statistics* **20**, 63–71.

Birch, M. W. (1964). The detection of partial association, I: the 2 × 2 case. *Journal of the Royal Statistical Society Series B* **26**, 313–324.

Breslow, N. E. and Cologne, J. (1986). Methods of estimation in log odds ratio regression models. *Biometrics* **42**, 949–954.

Breslow, N. E. and Day, N. E. (1980). Statistical Methods in Cancer Research I: *The Analysis of Case-control Studies.* Lyon: International Agency for Research on Cancer.

Hauck, W. W. (1984). A comparative study of conditional maximum likelihood estimation of a common odds ratio. *Biometrics* **40**, 1117–1123.

Holford, T. R., White, C., and Kelsey, J. L. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology* **107**, 245–256.

Lindsay, B., Clogg, C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86**, 96–107.

Lindsey, J. K. (2000). Directly modelling matched case-control data. *Statistics in Medicine* **19**, 35–44.

McCullagh P. and Nelder, J. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI* **22**, 719–748.

Mantel, N. and Hankey, W. (1975). The odds ratio of a 2 × 2 contingency table. *American Statistician* **29**, 143–145.

McCullagh, P. (1984). On the elimination of nuisance parameters in the proportional odds model. *Journal of the Royal Statistical Society Series B* **46**, 250–256.

Miettinen, O. S. (1970). Estimation of relative risk from individually matched series. *Biometrics* **26**, 75–86.

Miettinen, O. S. (1984). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine* New York: Wiley.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *The Canadian Journal of Statistics* **22**, 139–148.

Pike, M. C., Hill, A. P., and Smith, P. G. (1980). Bias and efficiency in logistic analyses of stratified case-control studies. *International Journal of Epidemiology* **9**, 89–95.

Platt, R. W., Leroux, B. G., and Breslow, N. (1999). Generalized linear mixed models for meta-analysis. *Statistics in Medicine* **18**, 643–654.

Rice, K. M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association* **99**, 510–522.

Robins, J., Breslow, N., and Greenland, S. (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**, 311–323.

Satten, G. A. and Kupper, L. L. (1990). Continued fraction representation for expected cell counts of a 2 × 2 table: a rapid and exact method for conditional maximum likelihood estimation. *Biometrics* **46**, 217–223.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics* **9**, 23–30.