

Equivalence of random-effects and conditional likelihoods for matched case-control studies

Ken Rice

MRC Biostatistics Unit, Cambridge, UK

January 8th 2004

Motivation

Study of genetic c-erbB-2 'exposure' and breast cancer (*Rohan et al, JNCI, 1998*)

		Number of cases exposed													
		0		1		0		1		0		1			
Number of controls exposed	0	1	0	0	2	1	8	2	15	1	12	0			
	1	0	0	1	3	0	1	0	6	1	6	1			
				2	0	0	2	1	1	3	1	2	3	0	
							3	0	0	3	1	0	3	0	0
										4	0	0	4	0	0
												5	0	0	

Each exposure measured imperfectly; Rate of False Positive Exposure ≈ 0.49

Rate of False Negative Exposure ≈ 0.00

(External validation study, 187 subjects)

Is any useful inference possible?

Matched case-control studies

- Some disease of interest, want to find if a binary 'exposure' is associated
- For each diseased case, find a control matched for other covariates; age, sex, etc
- *Then* measure exposure of interest
- The exposures are outcomes of interest, not the disease status
- must build a model for $\Pr(\text{exposure})$, not $\Pr(\text{disease})$
- Common study design, efficient, simple, popular

Formal description

- Control exposure (1 or 0) is Z_{1k} , Case exposure is Z_{2k} , for pair k
- $Z_{1k} \sim \text{Bern}(p_{1k})$, $Z_{2k} \sim \text{Bern}(p_{2k})$
- Assume odds ratio identical in all pairs k ;

$$\psi = \frac{p_{2k}}{1 - p_{2k}} \frac{1 - p_{1k}}{p_{1k}}$$

i.e. $\text{logit}(p_{2k}) = \log(\psi) + \text{logit}(p_{1k})$

- Generates one nuisance parameter for each pair

Problems!

- Maximum likelihood estimates are badly inconsistent
- Neyman-Scott problem – number of nuisance parameters grows with size of dataset
- Usual asymptotics not automatically valid
- ‘Sensible’ looking Bayes analysis can be even worse than MLEs!

Conditioning: a good solution

- Assume T_k = total number of exposures (0,1,2) doesn't contain information about ψ
- Condition on this (approx) ancillary statistic; conditional likelihood contributions are;

		Number of cases exposed	
		0	1
Number of controls exposed	0	1	$\frac{\psi}{1+\psi}$
	1	$\frac{1}{1+\psi}$	1

- Ratio of discordant pairs gives CMLE for ψ
- Well behaved, standard likelihood asymptotics work, but very hard to generalize

A wish list

- Analysis should reduce to conditional likelihood approach in standard situations
- Flexible method, easy to accommodate data which is less than ideal
- Allow use of prior information on ψ
- Fully model based, for simple interpretation
- Model criticism desirable, not currently well-supported

Random-effects: almost a dream solution

- All nuisance parameters, e.g. p_{2k} , drawn independently from G
Integrate likelihood w.r.t. p_{2k} , inference on ψ from marginal likelihood
- **Very** similar to a fully Bayesian approach;
 - mixing distribution $G \approx$ prior for p_{2k}
 - marginal likelihood \approx posterior for ψ (flat prior)
- Flexible, priors on ψ allowed, model based, model criticism possible
Just need to choose G – but no ‘default’ exists
- To complete the wish list, we need G which equate marginal and conditional likelihoods, if possible...

Random effects analysis

- Suppose $p_{2k} \sim G$, the mixing distribution
- Marginal likelihood contributions are;

		Number of cases exposed	
		0	1
Number of controls exposed	0	$1 \cdot E_G(\Pr(T = 0))$	$\frac{\psi}{1 + \psi} \cdot E_G(\Pr(T = 1))$
	1	$\frac{1}{1 + \psi} \cdot E_G(\Pr(T = 1))$	$1 \cdot E_G(\Pr(T = 2))$

- Define $E_G(\Pr(T = t)) = m_t$; the marginal probabilities

Equivalence

Lemma: Conditional likelihood = marginal likelihood
 if and only if
 G makes all m_t invariant with respect to ψ

- G which satisfy this are called **invariant** mixing distributions

Theorem: Invariant mixing distributions exist,
 for any matching ratio

- Lindsay *et al*, *JASA*, 1991, proved that for flexible G , CMLE and marginal MLE agree, but only for special datasets
- Invariant G depend on ψ
- Proofs follow by results on the Stieltjes Moment Problem

Invariant distributions: an example

- An example, for 1:1 matched case-control;

$$\begin{aligned}
 p_{1k} &= 1/2 && \text{with probability} && 1/2 \\
 p_{2k} &= 1/2 && \text{with probability} && 1/2
 \end{aligned}$$

- Dependence on ψ is present but implicit
- Nice ‘coin-tossing’ interpretation
- Get $m=\{0.25,0.5,0.25\}$ – other details about G don’t affect analysis
- Unchanged by relabelling case/controls, or exposure/non exposure; this property holds in some generality; is this ‘non-informative’?
- This example is ‘pretty’ but most aren’t!
Construction is essentially finding polynomial roots
Most applications just require existence – integrate over G to get m

Possible applications

These results allow us to put together the conditional analysis with the (many) benefits of a full likelihood approach;

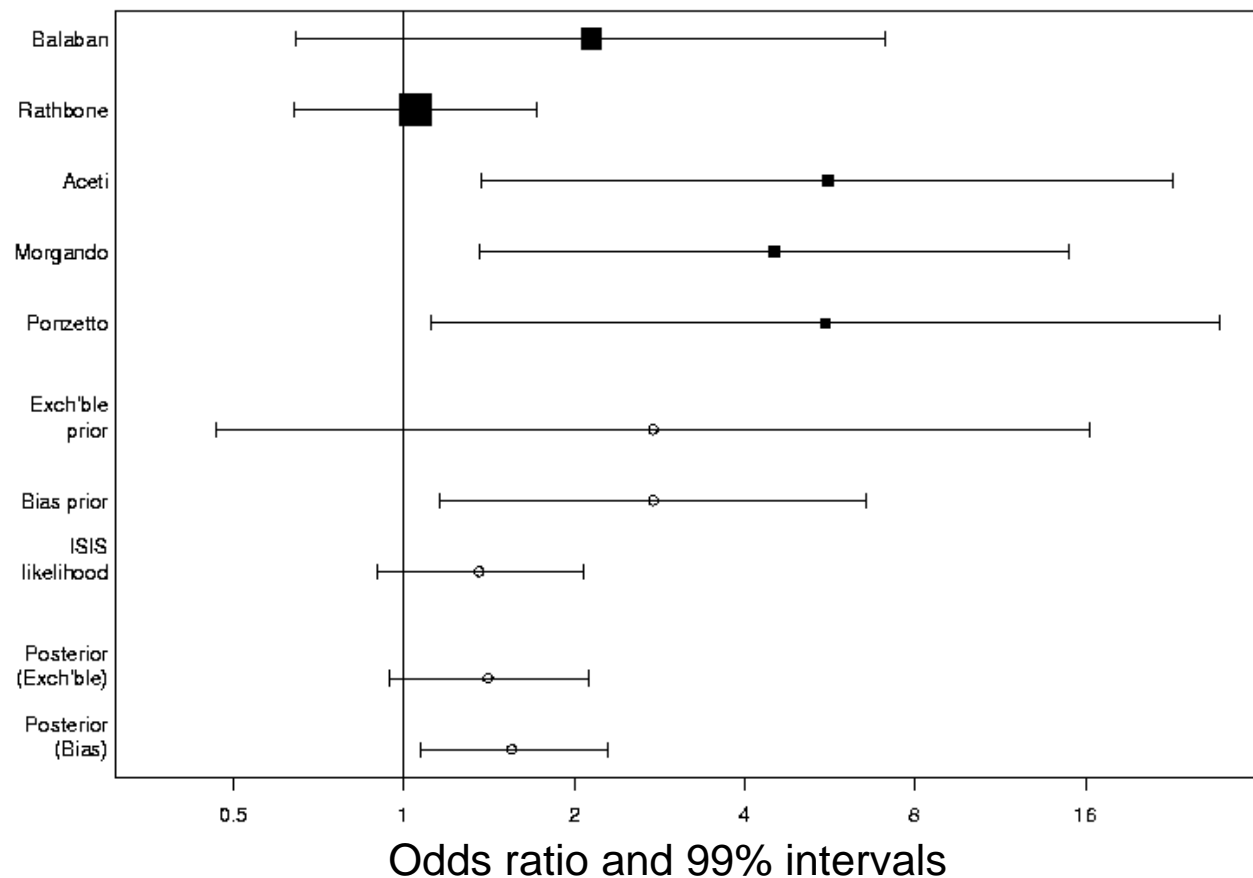
1. Using the conditional likelihood as a full likelihood;
 - Combining conditional analyses with prior information (ISIS)
 - No extra work

2. Fitting the conditional likelihood for ψ , and also fitting for m
 - Goodness of fit measures for conditional likelihood analyses (follows)
 - Allowing for misclassification in case-control studies (follows)
 - Inference on complex function of parameters, e.g. ranks in Rasch models
 - Involves complex likelihood function

3. MCMC algorithms for evaluating the conditional likelihood
 - Specify invariant distribution explicitly (polynomial roots)

Priors and conditional likelihoods

- Already used together, but necessary assumptions are now clear
- ISIS case-control study of helicobacter infection and myocardial infarction gave a 'ballpark' estimate incorporating prior beliefs - we can formalise this



Goodness of fit (1)

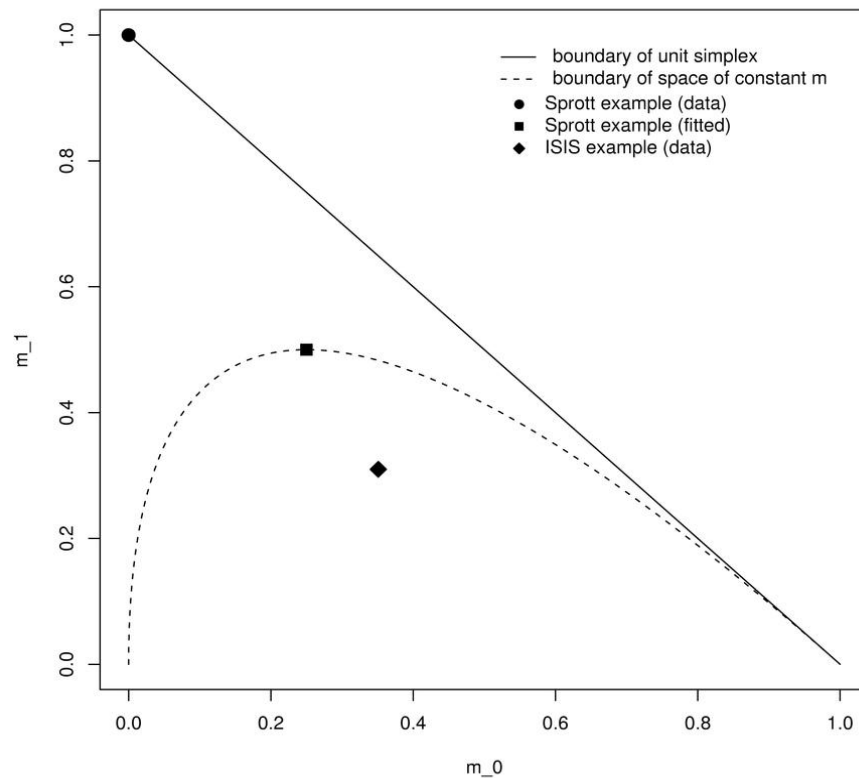
- A **binomial mixture** is a vector ν with elements which can be written

$$\nu_r = E_{F(\theta)} \binom{n}{r} \theta^r (1-\theta)^{n-r}, \quad r = 0, 1, \dots, n$$

- Previous $\{0.25, 0.5, 0.25\}$ corresponds to degenerate F ; $\theta = 1/2$ w.p. 1
- All m which correspond to invariant mixing distributions are binomial mixtures
- 1:1 correspondence holds in many (useful) special cases
- Leads directly to a measure of fit for the conditional ‘model’ – do the observed marginal totals T_k look like a binomial mixture?

Goodness of fit (2)

What does this space look like?



Sprott's example; conditional likelihood not appropriate

		Number of cases exposed	
		0	1
Number of controls exposed	0	0	50
	1	50	0

- Because m and ψ are orthogonal, some straightforward analyses aren't affected by this restriction

Misclassification

- Define X as the multinomial representation of Z_1, Z_2
- Usual measurement error model gives mixture for each data point;

$$\Pr(X = i) = \sum_j \Pr(X^* = j) \Pr(X = i | X^* = j)$$

- Assume 'true' data X^* from conditional 'model', in multinomial form
- Observed data X follow a multinomial model, although complicated by error probabilities
- Error probabilities can be known absolutely, or estimated
- Derived from sensitivity and specificity of exposure measurement

Return to the motivating problem

Study of genetic c-erbB-2 ‘exposure’ and breast cancer (*Rohan et al, JNCI, 1998*)

		Number of cases exposed													
		0 1		0 1		0 1		0 1		0 1					
Number of controls exposed	0	1	0	0	2	1	0	8	2	0	15	1	0	12	0
	1	0	0	1	3	0	1	1	0	1	6	1	1	6	1
				2	0	0	2	1	1	2	3	1	2	3	0
							3	0	0	3	1	0	3	0	0
										4	0	0	4	0	0
													5	0	0

$\Pr(\text{Observed Exposed} \mid \text{Unexposed}) \approx 0.49$

$\Pr(\text{Observed Unexposed} \mid \text{Exposed}) \approx 0.00$

(External validation study, 187 subjects)

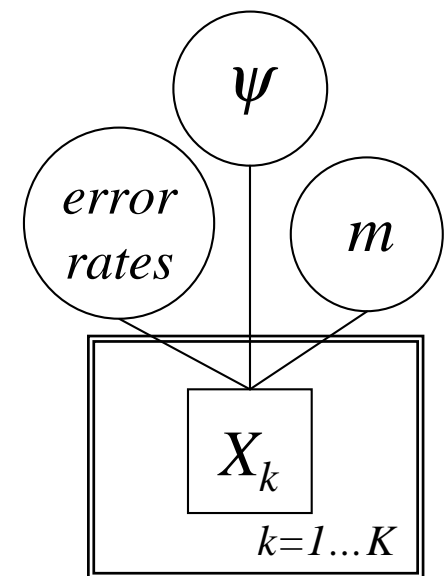
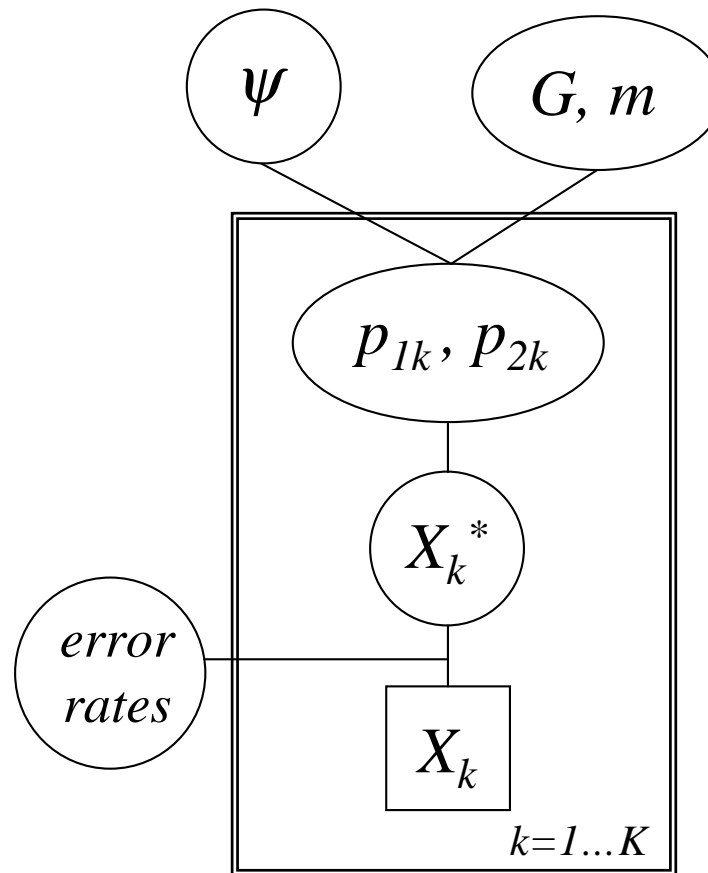
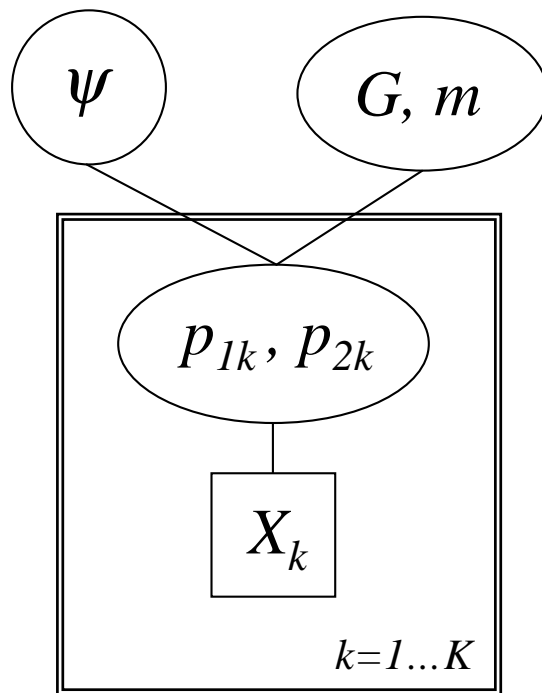
Assume:

Common odds ratio ψ

Invariant mixing distribution, with different vector m for 1:1, ... 1:5 matching

Extending the random effects model

- Perfect data approach
- Misclassified data
- Actual calculation



m constrained to be a binomial mixture

Application to breast cancer dataset

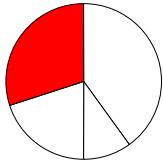
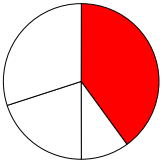
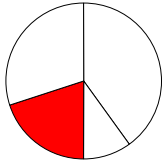
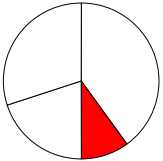
Analysis allowing for errors;

	Odds ratio estimate	False Positive Rate	False Negative Rate
Ignore errors in exposure	0.72 (0.30,1.69)	NA	NA
Use 'plug-in' error rates	0.66 (0.23,1.84)	0.49	0.00
With uncertain error rates	0.62 (0.17,1.68)	0.46 (0.34,0.59)	0.01 (0.00,0.04)

- Odds ratio estimate decreases, interval widens on the log scale (attenuation towards the null)
- Some inference *is* still possible, even with these error rates
- Simulations show intervals have good coverage (approx 95%)
- Estimates are slightly biased, on the log scale

Unusual estimate behavior (1)

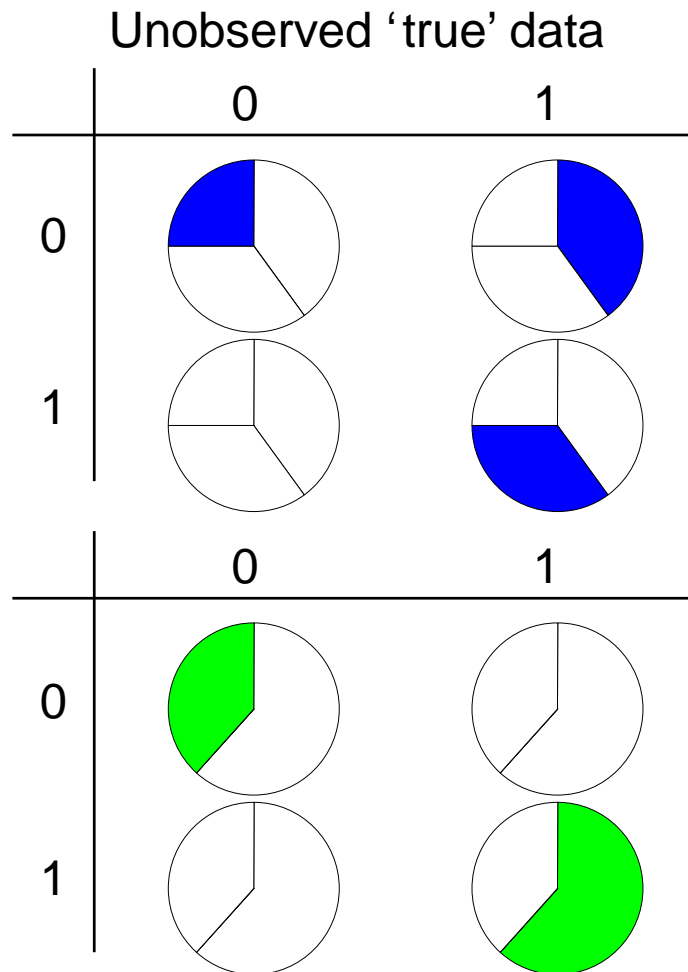
We do **not** impute the true data X^* 'above' our misclassified observations X , but the likely configurations characterise the analysis;

		Cell probabilities for 'true' data		Unobserved 'true' data	
		Number of cases exposed		0	1
Number of controls exposed	0	m_0	$\frac{\psi}{1+\psi} \cdot m_1$		
	1	$\frac{1}{1+\psi} \cdot m_1$	m_2		

- In this example, ratio of discordant pairs gives $\psi > 1$ and everything is 'nice'

Unusual estimate behavior (2)

Even with sensible data and error rates, ‘niceness’ is often absent;



- Most likely configuration is that all discordant pairs are same type
- The maximum likelihood estimate of ψ is at infinity
- Need confidence intervals which cope with extreme values of ψ
- Most likely configuration is that we have no discordant pairs
- The likelihood is maximized along a ridge; any value of ψ equally good
- Need a mechanism for reporting ‘no useful information’

The (simplified) ecological problem

- Assume a single 2x2 table,

	Exposed	Not exposed	Total
Controls	X_1	$n_1 - X_1$	n_1
Cases	X_2	$n_2 - X_2$	n_2
Total	$X_1 + X_2$	$n_1 + n_2 - X_1 - X_2$	$n_1 + n_2$

where only the marginal total $T = X_1 + X_2$ is observed

- Using an invariant prior for the nuisance parameter, the marginal likelihood for ψ is

~~$$L_{\text{conditional}}(\psi, X_1, X_2) \cdot m_T$$~~

- We **never** learn about ψ - can also occur with 'standard' priors, for special datasets

Do we again need to report 'no useful information'? Only if this model fits well?

Summary

- Conditional likelihood is a good approach for matched case-control studies
- An alternative derivation is available through random effects analysis
- The random effects derivation is easy to generalise and implement, allowing many new applications in matched case-control studies
- The random effects derivation uses the whole dataset, adding value to existing analysis at no 'cost' of more data, and providing new inferences in situations beyond matched case-control studies

Other ideas

- Rasch models; grid of binary outcomes,

$$\Pr_{i,j}(\textit{success}) = \frac{\alpha_i \beta_j}{1 + \alpha_i \beta_j}$$

	Q1	Q2	Q3	...
Student 1	0	1	1	...
Student 2	0	0	1	...
:	:	:	:	

Want to estimate ‘abilities’ α , condition out ‘difficulties’ β

- Categorical exposures;

Two nuisance parameters per pair

Two odds ratio parameters of interest

	<i>dd</i>	<i>dD</i>	<i>DD</i>
Case genotype	0	0	1
Control genotype	1	0	0

- Other non-standard likelihoods – Cox partial likelihood, already known to be approximately Bayesian; do the same ideas apply?
- Derivations of ‘good’ priors – our relabelling properties are not found in common non-informative priors; does this property guarantee ‘sensible’ analysis?

References and acknowledgements

Papers featuring work from this talk;

- Rice, K, Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, in press, *JASA*
- Rice, K, Discussion of Wakefield, J, 'Ecological inference for 2x2 tables', in press, *JRSS A*
- Rice, K and Holmans, P, Equivalence of conditional and marginal approaches to matched case control studies, with application to misclassification of a biallelic marker, in preparation
- Rice, K, Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies, *Statistics in Medicine*, 2003; **22**:3177-3194
- Duffy *et al*, Misclassification in a matched case-control study with variable matching ratio – application to a study of c-erbB-2 overexpression and breast cancer, *Statistics in Medicine*, 2003; **22**:2459-2468

Thanks to;

- Medical Research Council, ASA Epidemiology Section
- David Spiegelhalter, Stephen Duffy, David Clayton, Vern Farewell, Jon Wakefield and several anonymous referees