The risk ratio versus odds ratio argument revisited from a compositional data analysis perspective

J. BACON-SHONE

Social Sciences Research Centre, The University of Hong Kong, Hong Kong, johnbs@hku.hk

1. Introduction

Most statisticians, geologists and economists working with compositional data would now accept that the argument about whether to use log-ratios when modelling compositional data has been won by those following the approach of Aitchison (1985). It is clear that the Euclidean metric is not an appropriate measure for data in the simplex and should be replaced by the Aitchison distance. However, there are many other scientific disciplines that have not heard the message and continue to use distance metrics that are inappropriate for the sample space. In particular, when the composition is not directly observable, but is a vector of parameters in a model, there continues to be confusion as to what is an appropriate model. For example, the argument amongst epidemiologists and statisticians about whether to use risk ratios or odds ratios has raged for more than 30 years (Cummings, 2009). Those favouring risk ratios highlight the ease of interpretation by clinicians and that the risk ratio is not affected when adjustment is made by a variable that is not a confounder. Those favouring odds ratios point out that odds ratios are symmetrical with respect to both the outcome and risk variables, which is consistent with the likelihood ratio principle, unlike risk ratios.

The specific situation that I wish to examine here is when the composition is a set of (unobservable) probabilities. We will start with the very simple situation of two complementary probabilities representing the chance of success and failure commonly used in many epidemiological models. In these models, the aim is to understand the effect of a range of factors on the chance of death or survival.

2. A simple example

Consider the simplest situation, where we have a single factor X that is present or absent and we observe how many individuals with or without X present are alive or dead.

If p_1 is the probability of being alive with X present (labelled as 1) and p_0 is the corresponding probability when X is not present (labelled as 0) and n_1 and n_0 are the corresponding numbers of individuals found alive, while m_1 and m_0 are the corresponding numbers found dead, then the log likelihood function is clearly:

$$LL=n_1\log(p_1) + m_1\log(1-p_1) + n_0\log(p_0) + m_0\log(1-p_0)$$

In this situation, there is no problem finding a point estimate for the risk ratio p_1/p_0, for which the maximum likelihood estimate is:

$$n_1(n_0+m_0)/(n_0(n_1+m_1))\ ;$$

or for the odds ratio, which is $p_1(1-p_0)/(p_0(1-p_1))$, for which the maximum likelihood estimate is:

$n_1m_0/(m_1n_0)$.

Using likelihood ratio or Bayesian analysis, it is also straightforward to construct interval estimates for both measures.

Constructing exact confidence intervals for the risk ratio is difficult, although for the odds ratio it is straightforward after conditioning. As a result, it is common to use bootstrap intervals based on the maximum likelihood estimates.

3. The arguments

Cummings (1979) nicely summarizes most of the arguments between use of the risk ratio and odds ratio as follows:

3.1 Interpretation overall

The argument states that risk ratio is superior because it is more easily understood and used by clinicians. The supporting evidence is a long list of papers where the authors have clearly misunderstood the difference between risk ratio and odds ratio. In practice, if the risk is low under all scenarios, then there is little difference between the two measures. However, if the risk exceeds 10%, then the error in using the wrong measure is arguably significant.

3.2 Interpretability of averages

The estimate of the risk ratio (unlike odds ratio) will not change if we adjust for a variable that is not a confounder. This again means that it is easier to interpret a risk ratio, although if there are any confounders, this advantage disappears.

3.3 Constancy of odds ratios

As odds ratios are from R^+, it is possible for an odds ratio to be constant across a population, whereas because of the constraint on the risk ratio, this is impossible as there is an upper limit on probabilities, so the risk ratio cannot exceed the reciprocal of the unexposed risk. In short, risk ratio modelling does not address the implicit constraint. This argument should sound familiar to compositional data analysts and we will re-examine this argument in more detail below.

3.4 Symmetry of odds ratios

When calculating odds ratios, it makes no difference how we label outcome, we obtain the same result. However, risk ratios change if we change the labelling.

3.5 Estimation problems with risk ratios

Cummings (1979) does not discuss this problem, but Williamson (2011) devotes his entire thesis to discussing in detail how to address the problem that maximum likelihood methods often fail to converge for log binomial models, which involve fitting binomial models with linear models for the log risk ratio, which are the logical extension of the simple model considered above. He shows that for many methods commonly used for modelling risk ratios, there are problems with estimation using standard iterative methods because the likelihood maximum is on the boundary, as a consequence of the constraint on the probabilities.

4. More complex situations

For more complex scenarios, it is common to frame the problem in terms of generalized linear models, popularized in the book by Nelder and McCullagh (1989). In this formulation, the underlying distribution is that the outcome variables follow independent Bernoulli distributions with probability p_i of success, with a linear predictor

$\mu = X\ \beta$

where $\mu = g(p)$

for some link function g() that is a monotonic differentiable function.

The most common approach is to use the canonical link, which ensures that the linear predictor μ yields X^TY as the sufficient statistic, which in this case means using $g(p)=\log(p/(1-p))$.

Other possibilities considered by Nelder and McCullagh for the case of binary data are that g(p) is

$\log(-\log(1-p))$

$-\log(-\log(p))$

or

$\Phi^{-1}(p)$.

The models being used by those modelling risk ratios directly, however, correspond to g(p) is

$\log(p)$.

5. Discussion

What then does compositional data analysis have to offer, in helping to identify a resolution for this longstanding argument?

First, it will be recalled that the early arguments against log ratio analysis included claims that it must be faulty because it was harder to interpret than linear model analysis (or than mappings onto the sphere rather than R^d). It was only when people understood that some questions do not make sense for compositional data, such as trying to model the difference in compositions using Euclidean distance that progress could be made.

Second, we now recognise the need to either map all our questions onto R^d so that we can use the Euclidean metric or alternatively define a distance metric that is appropriate for the original sample space.

Let us now re-examine the difference between log binomial regression (models for risk ratios) and logistic regression (models for odds ratios)

As noted above, the standard log binomial linear model is based on the link function:

$g(p) = \log(p)$

Whereas for logistic regression, the matching link function is:

$g(p)=\log(p/(1-p))$.

On the face of it, there may seem no theoretical reason why one model should be superior to the other.

However, p and 1-p comprise a simple composition, even if p is a parameter, rather than data. Hence, it makes sense to a compositional data analyst that we must either use logistic regression with a Euclidean metric or equivalently model p on the simplex using the Aitchison metric.

If we examine the log binomial model, there is an obvious mapping problem because log(p) only covers R^-, so there is an implicit constraint on $X\beta$ to cover only R^- as well. This constraint is very problematic because as a constraint on β, it depends on X. This means that if we do two experiments with different X, then the constraints on β are different, so there is no simple way to combine our results. This also means that regardless of sample size, adding one new observation with a different value of X can significantly change the constraint for β and hence the estimate.

It is interesting to note that Nelder and McCullagh mention that all their suggestions for g(p) (unlike g(p)=log(p)) correspond to "inverses of well-known cumulative distribution functions having support on the entire real axis", although they do not explain why this matters, presumably because they consider this so glaringly obvious. They also go on to explain that the logistic function has the key advantage over other link functions of giving the same answer for prospective and retrospective sampling (i.e. conditioning on either row or column totals).

In short, it is impossible to frame sensible questions about β for a log binomial model unless there are fixed boundaries for X. In other words, we not only must know the sample space for Y, but must also have a finite sample space for X and cannot generate sensible models if the sample space for X is continuous.

This is clearly a very serious weakness of log binomial models. The advantage of risk ratios being easier to interpret is far outweighed by the difficulty of interpreting the parameters of the underlying models other than for very simple situations. Constructing linear models with constraints on the parameter space that depend on X does not seem sensible.

6. Conclusions

By reviewing the arguments about risk ratio versus odds ratio from a compositional data analysis perspective, it is clear that the log binomial models that underpin risk ratio models have serious flaws and cannot make sense if the sample space for X is not finite. In comparison, the logistic regression models that underpin odds ratio models are consistent with compositional data analysis principles and can handle any sample space for X, even if the consequence is models that may appear harder to interpret for clinicians.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and J. Bacon-Shone (1981). Bayesian risk ratio analysis. American Statistician: 254-257.

Cummings, P. (2009). The relative merits of risk ratios and odds ratios. Arch. Pediatr. Adoles. Med. 163 (5), 438-445.

McCullagh, P. and J.A. Nelder. (1989). Generalized Linear Models (second edition), Chapman and Hall, London.

Williamson, T.S. (2011). Log-Binomial Models: Maximum Likelihood and Failed Convergence, PhD thesis, University of Calgary, Calgary.