16 Case control studies

Theoretical basis for "odds ratio" as estimator of Rate Ratio, together with statistical model for the estimator

The old-fashioned and very loose justification for using the empirical odds ratio, *or*[1], as an estimator of the theoretical rate ratio goes back to Cornfield in the 1950s. Unfortunately it still is the one given in many 'modern' texts, despite the much more general justification provided by Miettinen in 1976.

The old justification rested on algebraic arguments using *persons*, not *population time*. The outcome *proportions* involved refer to *cumulative* incidence.

Clayton and Hills also start with proportions (risks), and thus are forced to use the now-very-old-fashioned "rare disease assumption". Later they argue that if we slice time very finely, we do not need this assumption, but it would be nice if they started with this modern way of viewing it.[2]

One way to be modern, and emphasize that we are generally in the 'rate' (rather than than 'risk') business is to use the empirical odds ratio as an estimate of the rate (i.e., intensity) ratio of interest, and to call it a rate ratio estimate; in other words, even if we derive the estimate from a crossproduct that *looks like* an odds ratio, or from the output of a logistic regression where the estimate is labelled 'odds ratio' (or the coefficient is labelled 'log odds ratio')

The truly modern way is to think of the cases as arising in population-time, and to think of the population time involved as an infinite number of person-moments - think of a person-moment as a person at a particular moment. Say that a proportion π_E of these are "exposed" person moments, and the remaining proportion π_0 are "non-exposed" person-moments. Suppose further that the (theoretical) event rates in the exposed and unexposed amounts of population-time are

$$\lambda_E = \frac{E[no.events]}{PT_E} \; ; \; \lambda_0 = \frac{E[no.events]}{PT_0},$$

with (theoretical) Rate Ratio $\theta = \lambda_E/\lambda_0$.

<u>Denominator Series</u> *[overall size d; d_0, d_1 in 'exposure' categories 0, 1]*

Suppose we take a finite random sample, of size d, of the infinite number of person moments in the base that generated the cases, and classify them into d_E "exposed" person moments and $d_0 = d - d_E$ "non-exposed" person-moments. We will refer to this sample of d as the *denominator* series. What is the statistical model for $d_E \mid d$? Clearly, it is

$$d_E \sim Binomial(d, \pi_E).$$

<u>Numerator (Case) Series</u> *[overall size c; c_0, c_1 in 'exposure' categories 0, 1]*

Denote by c the observed number of events; we classify them into c_E events in "exposed" population-time and $c_0 = c - c_E$ in the "non-exposed" population-time. We will refer to this sample of c as the *case* series.

What is the statistical model for $c_E \,|\, c$? We can think of c_E as the realization of a Poisson r.v. with mean (expectation) $\mu_E = (PT_E \times \pi_E) \times \lambda_E$. Likewise, think of for c_0 as the realization of a Poisson r.v. with mean (expectation) $\mu_0 = (PT_0 \times \pi_0) \times \lambda_0$.

Now, it is a statistical theorem (Casella and Berger, p194, exercise 4.15) that

$$c_E \mid c \sim Binomial(c, \mu_E/[\mu_E + \mu_0]).$$

Thus we can identify the distribution of the 4 random variables involved in the OR estimator

$$\hat{O}R = or = c_E/d_E \;\div\; c_0/d_0 \;=\; c_E/c_0 \;\div\; d_E/d_0 = (c_E \times d_0) \;\div\; (c_0 \times d_E).$$

The $c_E : c_0$ split is governed by one binomial, involving θ and other parameters, while the $d_E : d_0$ split is governed by a separate binomial, involving the same other parameters, but not involving θ.

[1] Sometimes I will use the lower case *or* to denote the observed or empiriacl odds ratio, and uppewr case *OR* to denote the true (but unobservable) odds ratio.

[2] There are a few instances where time *per se* isn't involved, e.g., success rates (with 'rate' as a proportion) of low versus high shots on ice hockey goalie Patrick Roy, acceptance rates of males versus females to medical school, etc. In these instances, it is still helpful to think of the 'successes' (or events) as numerators (classified as low/high, or male/female), and *samples* of <u>all</u> shots or applications, also classified in same way, as estimates of the relative sizes of the respective denominators. Note that if we sample from <u>all</u> candidates (shots or applicants) without excluding those that happen to form the numerator series, we *can directly estimate the ratio of the two proportions without ever involving anything that looks like an odds ratio.* To see this, imagine that candidates in the index and reference categories of the determinant in question would be expected to produce π_1 and π_0 'success' proportions respectively respectively, so that the estimand, the ratio of the two proportions, is π_1/π_0. Suppose we get to observe the two numerators y_1 and y_0, but not the full sizes N_1 and N_0 – the two denominators – of the two candidate pools. We can estimate the relative sizes by sampling say n from the $N_1 + N_0$ and classifying them into n_1 and n_0 respectively. If we know the sampling fraction f, then we can estimate the two 'success' proportions as $\hat{\pi}_i = y_i/[n_i \times (1/f)]$ and estimate their difference or their ratio. However, in the ratio, the sampling fraction cancels out, leaving us with the estimator $\widehat{\pi_1 \div \pi_0} = (y_1/n_1) \div (y_0/n_0)$ that does not require knowledge of the sampling fraction. In the statistical model in this context, y_1 and y_0 are two binomials with unknown denominators, N_1 and N_0, while the n_1/n split is a hypergeometric random variable (that could be approximated by a binomial if the sampling fraction is small. Theoretically, or via simulation, one can show that $(y_1/n_1) \div (y_0/n_0)$ is median-unbiased for $\pi_1 \div \pi_0$.

If one replaces μ_E and μ_0 by their constituents, one can show that the odds that an unexposed person-moment in the series of $c + d$ represents a "case" is $c : d$, whereas the corresponding odds for an exposed person moment is $(\theta \times c) \; : \; d$.

In other words, in the dataset of $c + d$,

$$logit[Prob[case|0] = \log(c/d) = \beta_0 \; ;$$

$$logit[Prob[case|E] = \log(c/d) + \log\theta = \beta_0 + \beta_E E,$$

where E is an indicator variable.

So, one can estimate $\log\theta = \log OR$ by a logistic regression of the Y's ($c + d$ observations in all, with $Y = 1$ if in case series; $= 0$ if in denominator series) on the corresponding set of $c + d$ indicators of exposure (1 if exposed, 0 if not).

C & H preamble

"If there is no relationship between exposure and disease incidence the distribution of exposure among the cases should be the same as the distribution among the controls."

We often use this reasoning informally, and supply a 'denominator series' from our own experience, as when we are confronted with the statement that 90% of all driving accidents occur within 10 Kilometres of home, or that in most pancreas cancer cases in North America the patient gives a history of coffee-drinking, or that most goals in ice hockey are on low shots, or that those in professional sports tended to be born in certain months of the year. We do not formally call this experience the 'control series'; rather it is more aptly referred to as a 'denominator' series.

Supplementary Exercise 16.1

Refer to the article "Road Trauma in Teenage Male Youth with Childhood Disruptive Behavior Disorders: A Population Based Analysis" by Redelmeier et al. in PLoS Medicine, November 2010, Volume 7, Issue 11, e1000369.

The following is an excerpt from the Abstract:

> A history of disruptive behavior disorders was significantly more frequent among trauma patients than controls (767 of 3,421 versus 664 of 3,812), equal to a one-third increase in the relative risk of road trauma (odds ratio = 1.37, 95% confidence interval 1.22–1.54, p,0.001). The risk was evident over a range of settings and after adjustment for measured confounders (odds ratio 1.38, 95% confidence interval 1.21–1.56, p¡0.001). The risk explained about one-in-20 crashes, was apparent years before the event, extended to those who died, and persisted among those involved as pedestrians.

In the Methods, the authors state:

> We excluded teenage girls from both groups to avoid Simpsons paradox (a spurious association created by loading on a null-null position) since this group has much lower rates of crash involvement.

In the Discussion, the authors state:

> A third limitation that causes our study to underestimate the association of disruptive behavior disorders with road trauma is that the data excluded girls [74]. To address this issue we retrieved the original databases, replicated our methods in girls rather than boys, and conducted a post hoc analysis. As anticipated, the results yielded a smaller sample (n = 4,156) and about the same estimated risk (odds ratio 1.31, 95% confidence interval 1.07–1.61, chi- square = 6.8, p = 0.010). Hence, the association of disruptive behavioral disorders with road trauma extended to both teenage boys and girls. Of course, many issues remain for future research including medication level at time of injury, amount of driving, extent of brain trauma, and sequelae among those not hospitalized [75,76].

Questions:

i. Reproduce the (crude) odds ratio and CI reported in the abstract.

ii. Figure out how the authors came up with the statement that "the risk [factor] explained about one-in-20 crashes".

Hint: (i) in what fraction of crashes was the factor present? [note that the factor cannot explain cases among those in in which the factor was absent].

(ii) Even when it was present, the factor wasn't responsible for all of these crashes. Most of them would have happened even in the absence of the factor. Use the reported rate ratio (take the 'adjusted' 1.38 rather than the crude 1.37). Work out, using say 138 accidents in males in whom the factor *was* present, what fraction of them would have occurred because of unrelated background factors and what fraction would be 'excess' cases due to the factor itself (or ask yourself: for every crash among those with the factor that was because of background (unrelated) causes, how many would there because of the factor? The (relative) rates in those with and without the factor are the key here). This latter fraction is called the 'etiologic fraction among the exposed'.

Then multiply this etiologic fraction (of cases among the exposed that are due to the exposure) by the fraction in (i) to get the overall fraction of all cases that is due to the factor.

This overall fraction (a fraction of a fraction) is called the overall or population etiologic fraction (sometimes called the population 'attributable' faction). The two formulae for it are not well understood by epidemiologists. See the article 'A heuristic approach to the formulae for the population attributable fraction' under the 'reprints' tab on JH's homepage.

iii. Come up with a reasonably realistic example of a behaviour/trait and the occurrence of some health event, where, in a 'case-control' study, if one simply added the 4 frequencies in the 2×2 table for boys and the corresponding ones for girls, and used the combined frequencies from this overall 2×2 table to produce one odds ratio, one could produce a very different odds ratio than the (say) common odds ratio in each of the gender-specific tables.

Do you think Redelmeier need have been concerned with Simpson's paradox in his context?

iv. The reported odds ratio for girls (1.31) is accompanied by both a confidence interval and a (null) chi-quare statistic and p-value. The CI was probably arrived at using Woolf's formula. Compute a test-based confidence interval instead and comment on how close it comes to the reported interval.

v. From the reported odds ratio of 1.31, and assuming that appendicitis is just about as common in boys and girls, that the 4,156 is the total number of trauma and appendicitis admissions in girls, and stating any other assumptions you are forced to make, try to reconstruct what the 2×2 table must have looked like for girls. (the reported CI should be of considerable help!)

vi. We briefly discussed in class how one could merge(combine) the odds ratios for boys and girls to get a single point estimate and associated CI. [notice that *combining the odds ratios* to get 1 new odds ratio can yield a very different result from *combining the raw frequencies* into one 2×2 table, and making one odds ratio from this one table]. Formally merge the results in 3 ways:

(a) Using the antilog of the weighted average of the logs of the gender-specific odds ratios (also known as Woolf's method) As part of this exercise, prove that the linear combination Woolf uses is the linear combination with the minimum variance.

(b) Using the Mantel-Haenszel summary odds ratio, and accompanying your point estimate by a *test-based* CI. [Whereas the M-H point estimate and test-statistic formulae date from 1959, we had to wait much longer for a specialized variance formula to accompany the point estimate.]

(c) Using a likelihood-based approach to estimation of $\theta = \log OR$, in which you represent each of the two items of data as normal-based log likelihoods centered on $\hat{\theta}_M$ and $\hat{\theta}_M$,then add the two log-likelihoods. Hint: since each log-likelihood is a quadratic form in θ, and since their sum is again a a quadratic form in θ, this amounts to working out where the new log-likelihood is centered, and what its curvature is. Show that its centre has the same form as the one used by Woolf.

17 Likelihoods for the odds ratio

Supplementary Exercise 17.1

Using the same types of calculations we did in R to replicate the results obtained by Fisher in his example comparing mono and dizygotic twins, replicated the calculations in section 17.4 of Clayton and Hills.