

Odds ratio, relative risk and case-control studies

Bob Wheeler
ECHIP, Inc.
bwheeler@gmail.com

1 Introduction

Many studies in the social sciences make use of the odds ratio to summarize data in 2x2 tables. In most cases, the odds ratio is used as an approximation to a statistic of greater interest, the relative risk, which is not always calculable because of the method of sampling. The justification for this is that they are close when the underlying risks are small, which is true; but since in the usual course of research this is not verifiable and since the difference in the risks is actually more important, odds ratios are reported that are sometimes quite different than the relative risk. In any case, contrary to the widespread belief, relative risk may be calculated in many studies without the need to use the odds ratio as an approximation.

2 2x2 tables

The observed odds ratio t is a statistic that is used to measure the disparity of values in a 2x2 table such as in Table 1.

	Treatment	No Treatment	Total
Result	a	b	$a + b$
No Result	c	d	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$

Table 1: Observed 2x2 table

It is calculated by dividing the observed odds in the first column $o_1 = (a/c)$ by the observed odds in the second column $o_2 = (b/d)$, which gives $o = o_1/o_2 = (ad)/(bc)$. It is important to note that o is symmetric and can equally well be taken as the observed odds ratios for the two rows. In addition, it is unaffected by multiplying rows or columns by a constant; such a multiplication corresponds to changing the amount of data taken for a given row or column, other things being equal. One may also calculate the observed risks and observed relative risk for the columns or rows; for columns, one has observed risks $r_1 = a/(a + c)$ and $r_2 = b/(b + d)$ and the observed relative risk $r = r_1/r_2$.

The assumption is that the observed values are a random sample of N observations from a table involving probabilities as in Table 2.

	Treatment	No Treatment
Response	$p_{1,1}$	$p_{1,2}$
No Response	$p_{2,1}$	$p_{2,2}$

Table 2: Probabilities 2x2 table

The probabilities in Table 2 sum to unity, and each observation is allocated at random to a cell in the table with respect to the probabilities. For this table the odds ratio is $(p_{1,1}p_{2,2})/(p_{1,2}p_{2,1})$, and the column risks are $p = p_{1,1}/(p_{1,1} + p_{1,2})$ and $q = p_{1,2}/(p_{1,2} + p_{2,2})$, as shown in Table 3. These risks are actually conditional probabilities: p is the conditional probability of a Response given Treatment, or $p = P(R|T)$ with R and T symbolizing Response and Treatment.

	Treatment	No Treatment
Response	p	q
No Response	$1 - p$	$1 - q$

Table 3: Column Risks 2x2 table

Data may be taken as in Table 2 or it may be taken conditionally with respect to the columns as in Table 4, or with respect to rows according to a similar table. For column data, the subjects are divided at random into two groups and one group is treated and the other not. Estimates of the odds ratio and relative risk may then be calculated as for Table 1. These risks are usually of most interest since they indicate how effective the treatment is with respect to the response.

A problem arises when the data is taken according to rows as in Table 4.

	Treatment	No Treatment
Case	r	$1 - r$
Control	s	$1 - s$

Table 4: Row Risks 2x2 table

The Cases row represents the risk for a sample of individuals who have responded and after the fact are categorized into Treatment and No Treatment groups. The Control row represents the risk for a sample of individuals who have not responded and are subsequently categorized into the two groups. Since the odds ratio is the same whether it is calculated by rows or columns, it may be used; but the column risks, which are of most interest, may not be calculated because the sample sizes in the two rows are arbitrary and simply doubling the sample in the second row changes the column risk estimates.

3 Relationship between odds ratio and relative risk

The odds ratio for Table 3 may be written $o = \rho \frac{1-q}{1-p}$, where $\rho = \frac{p}{q}$ which has led many to observe that o is very close to ρ when p and q are small, say less than 0.10. It is slightly more complicated than that, since the difference between the two also depends on the difference between p and q . A simple manipulation shows, $o = \rho(1 + \frac{\delta}{1-p})$, where $\delta = p - q$. This function is graphed in Figure 1. It may be seen that it departs substantially from unity for fairly small values of δ .

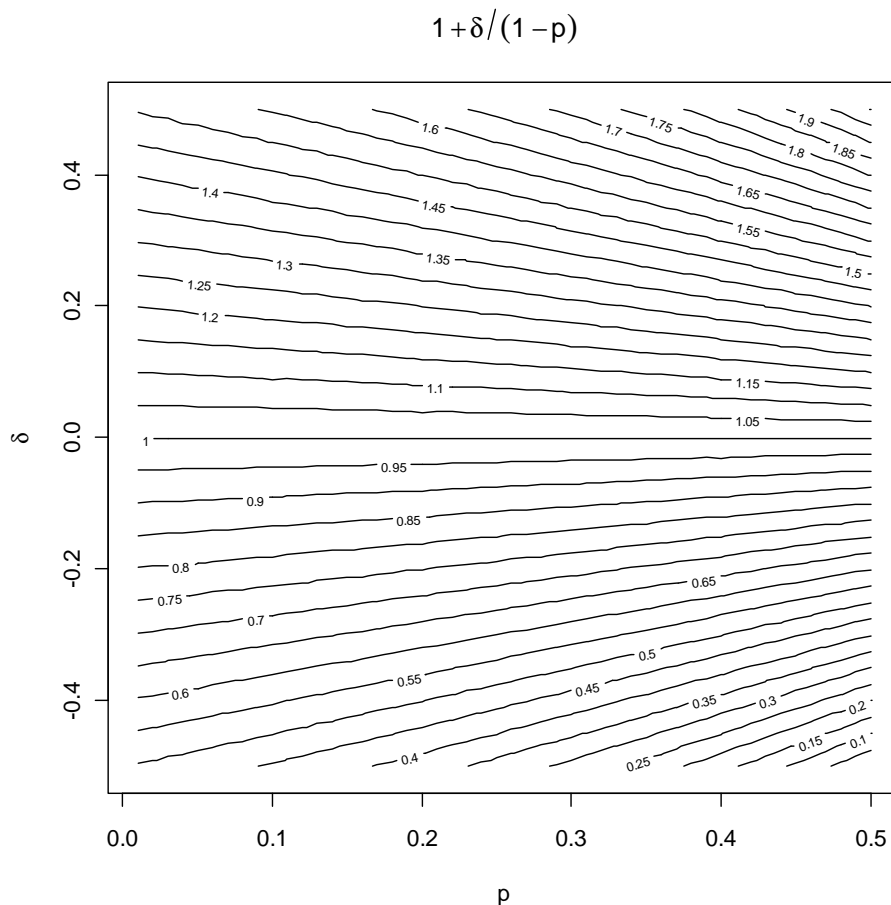


Figure 1: Relative risk multiplier

4 Estimating relative risk for case-control studies

It is not possible to estimate p from the data in Table 4 because r is a conditional probability with respect to Case: $r = P(T|C)$, where T is Treatment and C is case, and what is wanted is p the conditional probability with respect to Treatment: $p = P(C|T)$, which cannot be directly estimated because the data is taken according to rows and not columns.

If however the population probability $Q = P(C)$ of cases is known, then one has, *a la* Bayes.

$$P(T|C)Q = P(C|T)P(T),$$

$$p = P(C|T) = \frac{P(T|C)Q}{P(T)} = \frac{P(T|C)Q}{P(T|C)Q + P(T|Ct)(1-Q)},$$

where Ct is Control. In this equation, all probabilities have row wise estimates, except Q which is assumed known.

Thus upon substituting sample values from Table 1 and simplifying, one has the estimate,

$$\hat{p} = \frac{a}{a + cs} \text{ where } s = \frac{a + b(1 - Q)}{c + d - Q}.$$

This estimate, like the odds ratio, is unaffected by row or column scaling, and for non case-control studies, as illustrated by Table 1, $\frac{a+b}{c+d}$ approximates the population ratio $\frac{Q}{(1-Q)}$, and thus \hat{p} approximates the usual estimate $\frac{a}{a+b}$. The Bayesian relative risk is $\hat{r} = \frac{\hat{p}}{\hat{q}}$, where \hat{q} is defined similarly for the second column.

These Bayesian estimates are not overly sensitive to the value of Q . Table 5 shows simulated results for several values of Q around the true value of $Q = 0.2$. If a sample of 100 independent observations were taken to estimate Q , when the true value was 0.20, then the estimates would range from about 0.10 to about 0.30. As may be seen from Table 5 the resulting Bayesian estimates are reasonable. The odds ratio is 0.26, which is too low.

	True	Q=0.1	Q=0.2	Q=0.3
\hat{p}	0.10	0.05	0.10	0.16
\hat{q}	0.30	0.16	0.30	0.42
\hat{r}	0.33	0.29	0.33	0.38

Table 5: Various Q's

	$\hat{r}/\rho - \hat{o}/\rho$			
	q=0.01	q=0.05	q=0.10	q=0.30
p=0.01	1.03 - 1.02	1.00 - 0.96	1.00 - 0.95	0.99 - 0.70
p=0.05	0.99 - 1.00	0.98 - 0.98	1.00 - 0.95	0.99 - 0.70
p=0.10	1.00 - 1.10	1.01 - 1.07	1.00 - 1.03	1.01 - 0.79
p=0.30	1.00 - 1.42	1.00 - 1.36	1.00 - 1.29	1.00 - 1.00

Table 6: Relative Risks and Odds Ratios with respect to the true relative risk

A comparison of the relative risks and odds ratios for various small risk values is shown in Table 6. As may be seen, there is little to choose from for very small risk values, but the odds ratio is seriously degraded for larger values of the risks. Table 6 was obtained by simulation and the values in the table are geometric means.