BIOS 6244 Analysis of Categorical Data
October 5, 2005 Lecture

The Odds Ratio (Section 2.3, pp. 22-27)

Like the relative risk, the odds ratio (OR) also measures the association between exposure to some risk factor and a particular outcome (referred to generically as "disease"). Typically, the RR is used to measure the E-D association in prospective studies (e.g., cohort studies or clinical trials), whereas the OR is used in retrospective studies (e.g., case-control) and cross-sectional studies. The OR is also considered to be more robust than the RR as a measure of association, and there are certain types of statistical models (e.g., logistic regression) that can only be used with the OR. For retrospective studies in which the RR is not an appropriate measure of the E-D association, the OR can be used to estimate the RR as long as the disease under consideration can be thought of as "rare."

NOTE:
Throughout this lecture, "success" will be interchangeable with the outcome of interest (the "disease"), Population 1 will be interchangeable with the exposed population, and Population 2 will be interchangeable with the unexposed population.

As before, let π_1 = probability of "success" in Population 1, π_2 = probability of "success" in Population 2.

Then the *odds* of success in the 2 populations are defined to be:

$$odds_1 = \frac{\pi_1}{1-\pi_1} \quad \text{(Population 1)}$$

$$odds_2 = \frac{\pi_2}{1-\pi_2} \quad \text{(Population 2).}$$

That is, the *odds* of an event is defined to be the probability that the event will occur, divided by the probability that the event will not occur.

So, for example:

$\pi_1 = .75 \Rightarrow \text{odds}_1 = 3$

$\pi_1 = .5 \Rightarrow \text{odds}_1 = 1$

$\pi_1 = 0 \Rightarrow \text{odds}_1 = 0$

$\pi_1 = 1 \Rightarrow \text{odds}_1 = \infty.$

Note that for all events, $0 \leq \text{odds} < \infty$.

Odds can be easily converted to probability:

$$\pi_1 = \frac{odds_1}{odds_1 + 1}$$

and that, as a result, $\pi_1 = \pi_2$ if and only if $odds_1 = odds_2$.

The *odds ratio* for "success" in Population 1, relative to Population 2, is defined to be

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

Toxicology Example, cont.

Let π_1 = probability of developing a tumor if exposed to the natural chemical and π_2 = probability of developing a tumor if exposed to the synthetic chemical. Then

$OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ = odds ratio of developing a tumor if exposed to the natural chemical.

Properties of the Odds Ratio (Sec. 2.3.1)

Recall the terminology and notation that we introduced in September 28's lecture for 2x2 contingency tables.

Let X = the row variable and Y = the column variable. Then X and Y are independent if and only if $\pi_1 = \pi_2$, where π_1 = probability of "success" in Row 1, and π_2 = probability of "success" in Row 2.

We know that $\pi_1 = \pi_2$ if and only if $odds_1 = odds_2$ and OR = 1; thus, X and Y are independent if and only if OR = 1. Therefore, OR =1 is considered to be the null value in Epi studies of the E-D association, i.e., OR = 1 indicates that no E-D association is present.

Note that $0 \leq OR < \infty$. If OR > 1, the odds of "success" are greater in Row 1 than in Row 2. If OR < 1, the odds of "success" are less in Row 1 than in Row 2.

Two values of the OR represent the same strength of association, but in opposite directions, when one value is the reciprocal of the other.

Example

If OR = .25, then the association is considered to be just as strong as if OR = 4, but the association is in the opposite direction. That is, when OR = .25, Row 1 is associated with a *decrease* in risk, whereas when OR = 4, Row 1 is associated with a *increase* in risk.

When the order of the rows (or columns) of a 2x2 table is reversed, the new value of the OR is the reciprocal of the original value. Ordering of rows and columns in a 2x2 table is generally arbitrary, so whether we get OR = .25 or OR = 4 is simply a matter of how we label the rows and columns.

However, in Epi studies, we generally label the rows and columns as follows:

		Disease	
		Present	Absent
Exposure	Present		
	Absent		

Sometimes this is called the "canonical form" of a 2x2 table in studies of the E-D association.

Using this arrangement, π_1 = probability of disease in the exposed group and π_2 = probability of disease in the non-exposed group and the OR is referred to as "the odds ratio of disease, given exposure."

If the rows and columns are arranged in this way, then an exposure that increases the risk of disease will have OR > 1. An exposure that decreases the risk of disease will have OR < 1 and an exposure that is unrelated to the disease will have OR = 1.

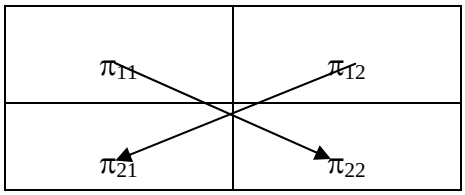
Note that the value of the OR is unchanged if we transpose the 2x2 table (i.e., swap rows and columns). In some instances, it is beneficial to present the 2x2 table in this alternative form. This illustrates one of the advantages of the OR over the RR: the value of the RR *does* depend on how we label rows and columns, whereas the value of the OR does not.

Using the notation previously defined for 2x2 tables, we can define the OR in terms of joint probabilities:

$$OR = \frac{\pi_{11} / \pi_{12}}{\pi_{21} / \pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The latter expression is called the *cross-product ratio* since it is the product

of the main diagonal entries, divided by the product of the off-diagonal entries:



To estimate the OR using sample data, we use sample proportions in place of population proportions:

$$\widehat{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

It is an exercise for you to show that

$$\widehat{OR} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

For the binomial sampling schemes that we described in the lecture on September 28, $\widehat{\text{OR}}$ is the MLE of the true OR.

Odds Ratios for the Physician's Aspirin Study (Sec. 2.3.2)

Consider Table 2.3, p. 20, in our text:

Table 2.3 Cross Classification of Aspirin Use and Myocardial Infarction (MI)

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *N. Engl. J. Med., 318*: 262–264 (1988).

For the physicians taking placebo,

$$odds\ of\ MI = odds_1 = \frac{n_{11}}{n_{12}} = \frac{189}{10845} = .0174,$$

i.e., there were 1.74 "yes" responses for every 100 "no" responses in this group.

For the physicians taking aspirin,

$$odds\ of\ MI = odds_1 = \frac{n_{21}}{n_{22}} = \frac{104}{10933} = .0095,$$

i.e., there were .95 "yes" responses for every 100 "no" responses in this group.

Then

$$\widehat{OR} = \frac{odds_1}{odds_2} = \frac{.0174}{.0095} = 1.832.$$

It is easier to use the computational formula:

$$\widehat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{(189)(10933)}{(104)(10845)} = 1.832.$$

Interpretation: The odds of MI were 1.83 times higher in the placebo group than in the aspirin group.

Alternatively, if we treat aspirin as the "exposure" (which would be more common, since it is the "experimental" exposure), we find OR = 1/1.832 = .546.

Interpretation: The odds of MI were .55 times as low in the placebo group as in the aspirin group.

This latter finding indicates that aspirin has a *protective effect* with regard to MI.

Statistical Inference for the Odds Ratio (Section 2.3.3)

Unless the sample size is quite large, the sampling distribution of $\widehat{OR}$ tends to be highly skewed. For example, if OR = 1, then $\widehat{OR}$ cannot be much smaller than OR since its lower bound is 0. However, $\widehat{OR}$ could be much larger than OR since it has no upper bound.

Since log transformations generally help with making highly skewed distributions more symmetrical (and also more like the normal distribution), traditional statistical inference for the odds ratio has focused on using $\log(\widehat{OR})$ to estimate log(OR).

Note that OR = 1 if and only if log(OR) = 0.

log(OR) is symmetric about 0, in the sense that interchanging rows and columns changes its sign. So, 2 values of log(OR) that are the same exept for sign, such as log(2) = .7 and log (.5) = -.7 represent the same strength of E-D association.

With regard to the sampling distribution of log($\widehat{OR}$), it can be shown that it is asymptotically normal with approximate mean and standard error given by:

$$E[\log(\widehat{OR})] \approx \log(OR) \text{ and}$$

$$SE[\log(\widehat{OR})] \approx \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}. \qquad (3)$$

The latter expression is sometimes referred to as the ASE (*asymptotic standard error*.)

Note that the ASE is infinite if any of the cell frequencies is 0.

By using our general approach for finding an approximate 95% CI(θ) for any parameter θ, we can find an approximate 95% CI[log(OR)], by using

$$\log(\widehat{OR}) \pm 1.96 ASE[\log(\widehat{OR})].$$

To find an approximate 95% CI(OR), we "back transform" the endpoints of this interval using exponentiation.

Example (Physician's Aspirin Study), cont.

log($\widehat{OR}$) = log(1.832) = .605

$$ASE[\log(\widehat{OR})] = \sqrt{\frac{1}{189}+\frac{1}{10845}+\frac{1}{104}+\frac{1}{10933}} = .123$$

So, an approximate 95% CI [log(OR)] = .605 ± 1.96(.123) = .605 ± .241 = (.364, .846) and an approximate 95% CI [(OR)] = ($e^{.364}$, $e^{.846}$) = (1.44, 2.33).

Since $1 \notin (1.44, 2.33)$, we reject H_0: OR = 1 and conclude that there is a significant reduction in 5-year risk of MI associated with taking aspirin.

Note that the approximate CI(OR) is not symmetric about $\widehat{OR}$, i.e., $\widehat{OR}$ = 1.83 is not the midpoint of the CI.

Zero Cells in a 2x2 Table

If any of the cells of a 2x2 table are 0, then either $\widehat{OR}$ =0 or $\widehat{OR}$ =∞, depending on whether the 0 cell occurs in the main diagonal or the off-diagonal, respectively. If 2 cells in a row or column are 0, then $\widehat{OR}$ is undefined. Furthermore, ASE[($\widehat{OR}$)] is infinite if any of the cells in the 2x2 table are zero.

How do we perform statistical inference using data like these?

Agresti recommends the use of an *amended* estimator:

$$\widehat{OR} = \frac{(n_{11}+.5)(n_{22}+.5)}{(n_{12}+.5)(n_{21}+.5)}$$

that can be used instead of the usual $\widehat{OR}$. In this case, ASE[log($\widehat{OR}$)] is calculated by adding .5 to each cell count in Equation (3) above.

Example
(Blondell RD, Dodds HN, Looney SW, Lewis CM, Hagan JL, Lukan JK, Servoss TJ, Toxicology screening results: Injury associations among hospitalized trauma patients. *Journal of Trauma*, 58, 561-570, 2005)

In a study of the association between positive tox screens and outcome among patients hospitalized with traumatic injuries, the following 2x2 table was obtained:

		Death	No Death
Cocaine	Positive	0	110
	Negative	30	739

Then $\widehat{OR}$ = 0 and ASE[log($\widehat{OR}$)] = ∞.

Using the amended table:

		Death	No Death
Cocaine	Positive	0.5	110.5
	Negative	30.5	739.5

we obtain:

$\widehat{OR}$ = .110, log($\widehat{OR}$) = -2.207, and ASE[log($\widehat{OR}$)] = 1.429.

So, an approximate 95% CI[log(OR)] is given by -2.207 ± 1.96(1.429) = .110 ± 2.801
= (-5.009, .595)

and an approximate 95% CI(OR) is given by ($e^{-5.009}$, $e^{.595}$) = (.01, 1.81).

There are also exact methods available for dealing with empty cells, which unfortunately are not available in either SAS or SPSS. However, LogXact is capable of performing the exact analysis. LogXact yields the following results:

$\widehat{OR}$ =0 with an exact 95% CI(OR) of (.00, .72).

Assignment 2: Due Monday, October 17