Case-Control Studies

Christopher Mecklin

March 29, 2016

So, I stated in the notes that we need to use the odds ratio OR and not the relative risk RR when we have a case-control study, but didn't elaborate on why this is the case. I was asked and bumbled through an answer, but here's a better answer.

Suppose Dr. Bradbury has developed a vaccine that is supposed to cure the Martian flu. Ideally, we would use a prospective (or longitudinal) study where we recruit subjects, randomly assign some to be vaccinated and others to receive a placebo, and then follow over time and ascertain how many from the vaccine group develop Martian flu and how many from the placebo group develop Martian flu. This might not be possible as it will take a lot of time and will require withholding the vaccine.

A *case-control study* is a study where the investigators pick cases (for example, people who have the Martian flu) and controls (people who do not have the Martian flu) rather than using randomization. In many situations this is necessary, as it is often either impossible or unethical to randomly assign people to the groups (i.e. randomly expose/infect people with the Martian flu).

Suppose we obtain the following data in our case control study. D^+ refers to subjects with the Martian flu (the cases), D^- to subjects without the flu (the controls) and T^+ subjects that received the vaccine.

	Vaccine T^+	No Vaccine T^-
Cases D^+	$a = 10$	$b = 30$
Controls D^-	$c = 100$	$d = 60$

The *odds ratio* OR is used to summarize the results from a case study. Among the cases, the odds of being vaccinated are 10/30 or 1/3. For the controls, the odds are 100/60 or 5/3. We then literally take the ratio of these odds.

$$OR = \frac{a/b}{c/d} = \frac{10/30}{100/60} = \frac{1}{5} = 0.2$$

When the odds ratio is less than one, we often take the reciprocal to make interpretation easier.

$$\frac{1}{OR} = \frac{5}{1} = 5$$

The odds of contracting the Martian flu are 5 times greater in the non-vaccination group than the vaccine group.

We will sometimes use the *log odds ratio* or *logit*. Notice that if the odds ratio is less than 1 (i.e. the cases are less likely to have the disease), the logit is negative. For example, $\ln 0.2$ = -1.609. If the odds ratio is exactly 1 (the cases and controls are equally likely to have the disease), the logit is zero: $\ln 0 = 0$. When the odds ratio is greater than 1, the logit is positive: $\ln 5 = 1.609$.

A 95% confidence interval of the log odds ratio is found with the following formula:

$$\ln OR \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\ln 5 \pm 1.96\sqrt{\frac{1}{10} + \frac{1}{30} + \frac{1}{100} + \frac{1}{60}}$$

$$1.609 \pm 0.784$$

The interval for the logit is $(0.825, 2.393)$. We then 'undo' the natural log transfomation to get a confidence interval for the odds ratio. This is similar to the process for finding a confidence interval for a correlation, where we computed the Fisher's z transfomation, found a CI for Fisher's z, and then 'undid' the transformation to yield an interval on the correlation.

We undo by exponentiating: $(e^{0.825}, e^{2.393})$ or (2.282,10.946). Notice the entire confidence interval lies above $OR = 1$, which serves as the 'null hypothesis' value. This indicates there is a significant relationship between the vaccine and contracting or not contracting the Martian flu.

This data can be analyzed with either Pearson's chi-square test or Fisher's exact test. The R output for Fisher's exact test gives an estimate and confidence interval for the odds ratio. The values are slightly different as the computer uses a more complex formula for these computations. Notice to get computer results based on the reciprocal of the odds ratio, that I switched the vaccine and no vaccine columns (this switch does not affect the calculation of either Pearson's chi-square or Fisher's exact test).

```
##          No Vaccine Vaccine
## Cases            30      10
## Controls         60     100
```

```
##          No Vaccine Vaccine Total Count
## Cases          75.0    25.0   100    40
## Controls       37.5    62.5   100   160
```

```
##
##  Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 18.182, df = 1, p-value = 2.008e-05
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  .Table
## p-value = 2.907e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   2.172895 12.216913
## sample estimates:
## odds ratio
##   4.958566
```

OK, so why can't we correctly use the relative risk. Well, we can compute it, but it is meaningless. Let's go ahead and compute it, then show why it is a useless statistic for a case-control study.

	Vaccine T^+	No Vaccine T^-
Cases D^+	$a = 10$	$b = 30$
Controls D^-	$c = 100$	$d = 60$

$$RR = \frac{a/(a+c)}{b/(b+d)} = \frac{10/110}{30/90} = \frac{3}{11} = 0.273$$

You could, of course, take the reciprocal.

But the number of cases and controls chosen is an arbitrary choice made by the researchers (sometimes driven by how easy or hard it is to find 'cases'). In my example, we had 40 cases and 160 controls, for 4 times as many controls. Suppose I have an equal number of cases and controls, with the same proportion of cases that were vaccinated or not vaccinated.

	Vaccine T^+	No Vaccine T^-
Cases D^+	$a = 40$	$b = 120$
Controls D^-	$c = 100$	$d = 60$

$$RR = \frac{a/(a+c)}{b/(b+d)} = \frac{40/140}{120/180} = \frac{3}{7} = 0.429$$

The relative risk is different. Does this happen to the odds ratio? Remember, the odds ratio of the original table (without taking reciprocal) was $OR = \frac{1}{5} = 0.2$.

$$OR = \frac{a/b}{c/d} = \frac{40/120}{100/60} = \frac{1}{5} = 0.2$$

It stays the same, and this is why the odds ratio is the choice for case-control studies!