Odds Ratio and Relative Risk

Collin Nolte

March 08, 2022

Population Stratification

Last week we considered the transmission disequilibrium test and made note that it controlled for effects of *population stratification*

As this will come up again in future lectures, it's worth briefly covering now. We'll start with a working definition

Population stratification is the presence of systematic differences in allele frequencies between subpopulations in a population as a results of *non-random mating* between individuals. As such, it is an important confounding variable in genome association studies

Here, we will consider two populations X and Y that together make up a larger population, Z

Population X	
AA	70
BB	10
AB	20

Population Y	
AA	30
BB	40
AB	30

Population Z	
AA	100
BB	50
AB	50

Because of Population X, A has a much higher allele frequency than B in the general population

The primary consequence of the higher allele frequency lies in the relative portions of homozygous parents, either AA or BB

Population Z	
AA	100
BB	50
AB	50

Transmitted	Non-transmitted		Total
	A	B	
A	100	b	
B	c	50	
Total			200

In disease association studies, we control for this confounding by only considering the offspring of those parents who are heterozygous

If there is no association between the disease and allele, we should expect that of the heterozygous parents, 50% would transmit A while 50% would transmit B

Population Z	
AA	100
BB	50
AB	50

Transmitted	Non-transmitted		Total
	A	B	
A	100	25	
B	25	50	
Total			200

The p-value for this table using McNemar's test is $p = 0.887$

If, on the other hand, there *is* an association between disease status and, say, allele A, we would expect over-representation of A passed from heterozygous parents

Population Z	
AA	100
BB	50
AB	50

Transmitted	Non-transmitted		Total
	A	B	
A	100	40	
B	10	50	
Total			200

The p-value for this table using McNemar's test is $p = 0.00004$

Review

Last week, we considered the χ^2 test, Fisher's Exact Test, and the exact binomial test for analyzing *association* of groups in 2×2 tables

Critically, we understand that these were *binary* tests of association, telling us nothing about the *magnitude* of association (or lack of association) between groups

Today, we will focus on methods for determining this magnitude

Odds

Odds are expressed as a proportion of successes to failures. For example, using a fair six sided die, we would say that the odds of rolling a six are 1 to 5 (one six to five not-sixes), or 0.2.

We could equally say the odds of not rolling a six to rolling a six are 5 to 1, or 5.0. In other words, depending on what is classified as success as failure in a binary outcome gives us two different odds that are inverses of each other

What we may be interested in comparing is the odds of some outcome between two populations

Odds Ratio

Treatment Status	Side Effects		Total
	Yes	No	
Control	10	40	50
Treatment	30	20	50
Total	40	60	100

For the control group, the odds of having side effects are 10 to 40, or 0.25. Similarly, the odds of having side effects for the treatment group is 30 to 20, or 1.5

From here, two equivalent statements can be made:

Treatment Status	Side Effects		Total	Odds
	Yes	No		
Control	10	40	50	0.25
Treatment	30	20	50	1.5
Total	40	60	100	

We might consider how the odds between Control and Treatment groups are related, and from here, two equivalent statements can be made:

Comparing the odds of side effects from treatment to control, we have an *odds ratio* of $1.5/.25 = 6$. That is, the odds of having side effects are 6 times higher in the treatment group, compared to control.

Equivalently, we also could have described the odds from control to treatment as $.25/1.5 = 0.1667$

Interpretation of Odds Ratio

Typically, we compute the odds in a standard way, which divides the odds of the first row by the odds of the second. Of course, which group is on which line is arbitrary

When $1 < \theta < \infty$, the odds of the first response are θ times higher in the first row than the second. When $0 < \theta < 1$, the odds are $(1/\theta)$ higher in the second row than the first. $\theta = 1$ *if and only if* the two groups are independent

The odds ratio remains the same if the columns and rows of the table are switched: in other words, θ is not determined by which margins (if any) are assumed to be fixed

$$OR = \hat{\theta} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Odds Ratio Test Statistic

Having an estimate for the odds ratio is nice, but otherwise meaningless without having a sense of the variability in this estimate. In other words, it would be convenient if we could compute a confidence interval as well.

Under the null $H_0 : \theta = 1$, we see that the odds ratio is just as likely to fall in the interval $(0, 1]$ as it is $[1, \infty)$. Consequently, the distribution of $\hat{\theta}$ is hardly normal. It can be shown that this is remedied with a log transformation:

$$\log \hat{\theta} \sim N(\log \theta, \hat{\sigma}(\log \hat{\theta}))$$

where

$$\hat{\sigma}(\log \hat{\theta}) = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)^{1/2}$$

Odds Ratio and CLT

Given that $\log \hat{\theta}$ is approximately normally distributed, we can compute the same style of confidence intervals that we used in the CLT. With $z_{\alpha/2}$ being our critical value, we can construct a $(1-\alpha)\%$ confidence interval as

$$\log \hat{\theta} \pm z_{\alpha/2}(\hat{\sigma}(\log \hat{\theta}))$$

Taking the exponential of the endpoints of this interval, we end up with a confidence interval for the odds ratio

$$\exp\left(\log \hat{\theta} \pm z_{\alpha/2}(\hat{\sigma}(\log \hat{\theta}))\right) = \exp\left(\log \hat{\theta}\right) \exp\left(\pm z_{\alpha/2}(\hat{\sigma}(\log \hat{\theta}))\right)$$

$$= \hat{\theta} \exp\left(\pm z_{\alpha/2}(\hat{\sigma}(\log \hat{\theta}))\right)$$

This estimate tends to be slightly conservative, leading to the possibility of incorrectly rejecting a null hypothesis

95 % CI for Odds Ratio

Treatment Status	Side Effects		Total	Odds
	Yes	No		
Control	10	40	50	0.25
Treatment	30	20	50	1.5
Total	40	60	100	

$$\hat{\theta} = 6, \quad \log\hat{\theta} = 1.792, \quad z_{\alpha/2} = 1.96$$

$$\hat{\sigma}(\log\hat{\theta}) = \left(\frac{1}{10} + \frac{1}{40} + \frac{1}{30} + \frac{1}{20}\right)^{1/2} = 0.456$$

Then

$$\log\hat{\theta} \pm z_{\alpha/2}(\hat{\sigma}(\log\hat{\theta})) = (0.8972, 2.6864)$$

$$\hat{\theta}\exp\left(\pm z_{\alpha/2}(\hat{\sigma}(\log\hat{\theta}))\right) = (2.4526, 14.6783)$$

R

A slightly more sophisticated version of this test is returned by the `fisher.test` in R

```
> fisher.test(mm)

        Fisher's Exact Test for Count Data

data:  mm
p-value = 0.000083
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  2.2658 16.3909
sample estimates:
odds ratio
    5.8813
```

Relative Risk

A more meaningful statistic for clinicians is often the *relative risk*, which expresses a ratio of probabilities of outcomes between populations based on exposure. For example, we might ask, "What is the proportional increase in heart disease for smokers compared to non-smokers?"

Caution needs to be used here as we are asking a question about conditional probabilities. Computing this directly requires knowing the probability of an outcome *in general*, which we may or may not have, depending on the study

RR

Population	Variable		Total
	+	−	
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	N

$$\pi_{1|1} = \{\text{Probability of '+' outcome, given Population A}\} = \frac{a}{a+b}$$

$$\pi_{1|2} = \{\text{Probability of '+' outcome, given Population B}\} = \frac{c}{c+d}$$

$$RR = \frac{\pi_{1|1}}{\pi_{1|2}} = \frac{a(c+d)}{c(a+b)}$$

RR Valid

A report published in 1988 summarizes results of a Harvard Medical School clinical trial determining effectiveness of asprin in preventing heart attacks in middle-aged male physicians

Treatment Status	Myocardial Infarction	
	Attack	No Attack
Placebo	189	10,845
Asprin	104	10,933

Here, the row totals were fixed by study design. As the column values were not determined a priori, these observations reflect the probabilities of an outcome in the study

RR Valid

Treatment Status	Myocardial Infarction	
	Attack	No Attack
Placebo	189	10,845
Asprin	104	10,933

$$\pi_{1|1} = \{\text{Probability of '+' outcome, given Population A}\} = \frac{189}{11034}$$

$$\pi_{1|2} = \{\text{Probability of '+' outcome, given Population B}\} = \frac{104}{11037}$$

$$RR = \frac{\pi_{1|1}}{\pi_{1|2}} = \frac{a(c+d)}{c(a+b)} = \frac{189 \cdot 11037}{104 \cdot 11039} = 1.812$$

RR Not Valid

A report published in 1950 summarizes the results of a case-control study between 20 London hospitals investigating relationship between cigarette smoking and cancer. The study involved 1,418 patients, matching 709 cases with lung cancer against 709 controls without lung cancer on the basis of gender and age

Smoking Status	Lung Cancer	
	Present	Not Present
Smoker	688	650
Non-Smoker	21	59
Total	709	709

Now it is the totals for the *outcome* that we determined a priori. From this data, it would be impossible to determine the actual prevelance of lung cancer within the population. Without these probabilities, relative risk cannot be computed

Odds Ratio vs Relative Risk

Population	Variable		Total
	+	−	
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	N

Fortunately, the odds ratio can be used as an approximation to the relative risk when incidence of the disease is close to zero (i.e., rare diseases). In such cases, relative to b and d, a and c will be small:

$$RR = \frac{a(c+d)}{c(a+b)} \approx \frac{ad}{bc} = OR$$

Comparison

Consider this in the Myocardial Infarction study, where the incidence of a heart attack is relatively small

Treatment Status	Myocardial Infarction	
	Attack	No Attack
Placebo	189	10,845
Asprin	104	10,933

$$RR = \frac{a(c+d)}{c(a+b)} = 1.812$$

$$OR = \frac{ad}{bc} = 1.832$$

Comparison

Returning to the smoking example, as the true incidence of lung cancer is known to be relatively small, the odds ratio can still provide a rough indication of relative risk

Smoking Status	Lung Cancer	
	Present	Not Present
Smoker	688	650
Non-Smoker	21	59
Total	709	709

$$OR = \frac{688 \cdot 59}{21 \cdot 650} = 2.974$$