Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 15, part 4: Inference for Odds Ratios

Jinko Graham

2018-10-18

Estimating Odds Ratios

- For Doll and Hill's lung-cancer data, we estimated the odds ratio from the sample proportions of smokers in the cancer (case) and non-cancer (control) groups.

		case	control
Smoke	Yes	$a = 1350$	$b = 1296$
(E)	No	$c = 7$	$d = 61$
		$a + c = 1357$	$b + d = 1357$

- OR estimate is

$$\widehat{OR} = \frac{\frac{a}{a+c}}{1 - \frac{a}{a+c}} \Bigg/ \frac{\frac{b}{b+d}}{1 - \frac{b}{b+d}} = \frac{ad}{bc}$$

- For Doll and Hill's data, we have

$$\widehat{OR} = \frac{ad}{bc} = \frac{1350 \times 61}{1296 \times 7} = 9.1.$$

Testing whether $OR = 1$

- The chi-square test assesses the null hypothesis that $OR = 1$ (no association between exposure and disease) against the alternative hypothesis that $OR \neq 1$ (an association).

```
##            case control
## smoker     1350    1296
## non-smoker    7      61
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mydf
## X-squared = 42.37, df = 1, p-value = 7.552e-11
```

- The p-value for testing $H_0 : OR = 1$ vs. $H_a : OR \neq 1$ is 7.552×10^{-11}.
- Strong evidence of association between lung cancer and smoking!
- Recall from the Chapter 6 notes that $OR \approx RR$, provided that the disease is rare.
- Assuming lung cancer is rare, we may approximate the relative risk of lung cancer by the odds ratio.
    - We estimated the OR to be $\widehat{OR} = \frac{a*d}{b*c} = \frac{1350*61}{1296*7} = 9.1$
    - So we estimate that the risk of lung cancer in smokers is about 9.1 times the risk of lung cancer in non-smokers (i.e. $\widehat{RR} = 9.1$).

Confidence Intervals for ORs

- Recall: The natural logarithm $\log_e(x)$ is defined so that $x = e^{\log_e(x)}$, where the base $e \approx 2.718$.
  - To get x, we **exponentiate** the natural logarithm $\log_e(x)$; i.e., we raise the base e to the power of the exponent $\log_e(x)$.
- For large samples, it turns out that $\log_e(\widehat{OR})$ is approximately normally distributed.
- This approximation leads to CIs for $\log_e(OR)$ of the form

$$\text{estimate } \pm m.e., \quad \text{where}$$

  the margin of error term, m.e., is an SE times a critical value.
  - The standard error of $\log_e(\widehat{OR})$ is:

$$SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

  - For a level-C CI for $\log_e(OR)$, the critical value is z^*, the upper $(1 - C)/2$ critical value of the standard normal distribution.

- A level-C CI for $\log_e(OR)$ is thus $\log_e(\widehat{OR}) \pm z^* SE$, or

$$\left(\log_e(\widehat{OR}) - z^* SE,\ \ \log_e(\widehat{OR}) + z^* SE\right),$$

- To get the level-C CI for OR, we exponeniate the lower and upper bounds of the CI above:

$$\left(e^{\log_e(\widehat{OR}) - z^* SE}, \quad e^{\log_e(\widehat{OR}) + z^* SE}\right)$$

Application to Doll and Hill's Lung-Cancer Data

- The estimated OR is $\widehat{OR} = \frac{1350*61}{1296*7} = 9.1$, and its logarithm is $\log_e(9.1) = 2.21$.
- The SE of $\log_e(\widehat{OR})$ is

$$SE = \sqrt{\frac{1}{1350} + \frac{1}{1296} + \frac{1}{7} + \frac{1}{61}} = 0.401$$

- For a 95% CI, the critical value is $z^* = 1.96$.
- The 95% CI for $\log_e(OR)$ is therefore

$$(2.21 - 1.96 \times 0.401, 2.21 + 1.96 \times 0.041) = (1.42, 3.00)$$

- The 95% CI for OR is then

$$\left(e^{1.42}, e^{3.00}\right) = (4.14, 20.1)$$

Interpretation of Point Estimates of OR

- Smoking is associated with an estimated 9.1-fold increase in the odds of lung cancer.
- Or, if lung cancer is **rare**, we can interpret the OR as an RR and say:
  - smoking is associated with an estimated 9.1-fold increase in the ***risk*** of lung cancer.

Interpretation of Interval Estimates of OR

- "With 95% confidence, smoking is associated with an estimated 4.1 to 20-fold increase in the odds of lung cancer"
  - You can use the statement above, but keep in mind that it really means that: In 95 out of 100 datasets, we expect the CI, such as 4.1-20 for this dataset, to cover the true OR.
- Assuming that lung cancer is rare, we can interpret the OR as an RR and say:
  - "With 95% confidence, smoking is associated with an estimated 4.1 to 20-fold increase in the *risk* of lung cancer"
  - i.e, in 95 out of 100 datasets, we expect the CI, such as 4.1-20 for this dataset, to cover the true RR.

More Than Two Exposure Levels

- Doll and Hill's data with smokers classified by the average number of cigarettes per day:

		case	control
Number of	25+	340	182
cigarettes	15-24	445	408
per day	1-14	565	706
	0	7	61

- Can use the last row with 0 cigs per day (unexposed) as a baseline group ($c=7, $d = 61$), and calculate our ORs for each level of exposure.
  - E.G. Estimated OR for 25+ vs. 0 cigs per day:

$$
\begin{aligned}
OR &= \frac{\text{odds of lung cancer in exposed (25+ cigs/day)}}{\text{odds of lung cancer in unexposed } (O \text{ cigs/day})} \\
&= \frac{a * d}{b * c} = \frac{340 * 61}{182 * 7} = 16.28
\end{aligned}
$$

Odds-ratios with Multiple Exposure Levels

- Add estimated *OR*s to the table:

		case	control	$\widehat{OR}$
Number of	25+	340	182	16.28
cigarettes	15-24	445	408	9.50
per day	1-14	565	706	6.97
	0	7	61	–

- The increase in estimated *OR*s with exposure level suggests a "dose-response" relationship.
- For observational data such as these, a dose-response relationship is one of the criteria for establishing causality.
  - In this study, the dose-response relationship with number of cigarettes per day was used to argue that smoking **causes** lung cancer.

Including Confidence Intervals

		case	control	$\widehat{OR}$	95% CI
Number of	25+	340	182	16.28	(7.30,36.32)
cigarettes	15-24	445	408	9.50	(4.30,21.02)
per day	1-14	565	706	6.97	(3.17,15.37)
	0	7	61	–	–

Historical notes

- Doll and Hill's study of lung cancer was published in 1954
  - Turned the tide of public-health opinion on smoking.
- Famously, the iconic geneticist and statistician RA Fisher was strongly opposed (to the point of campaigning against it for the tobacco industry).
- Fisher was a heavy pipe smoker
  - Argued that correlation (association) is not causation
- Died aged 72 in 1962, following complications from cancer surgery.

Chapter 15 Summary

- Contingency tables summarize the joint distribution of two categorical variables.
- The chi-square test tests for association between two categorical variables
- For data that are paired in some way, we use McNemar's test, which contrasts the discordant cells in a table that counts each pair just once.
- Testing for association in a 2×2 table amounts to testing whether or not the OR is 1.
  - Can extend to $r \times 2$ tables with r levels of the exposure.
- When the disease outcome is rare, the OR can be interpreted as a relative risk (RR).
- Can obtain approximate CIss for the OR.
- Omitted Berkson's fallacy in text; beyond scope of course.