14: Case-Control Studies

Independent Samples

Introduction

In the previous chapter we considered the basics of cohort analysis by comparing incidences (risks) of disease in two groups of people. Group 1 was "exposed" and Group 0 was not. The relationship between the exposure and disease was quantified in the form of a relative risk.

In this chapter, we consider the basics of case-control analysis. In case-control analysis, we use a disease-selective subset of the cohort to select cases and controls. Group 1 is the cases, and Group 0 is the controls. This precludes the possibility of direct estimation of incidence (risk), but retains the ability to estimate relative risk through an odds ratio. Justifications for this approach is complex, but basically relies on two possibly inter-related conceptions. Conception 1 is based on a Baysian proof,[1] whereas conception 2 is based on density sampling of the population at risk.[2] Both justifications are worthy of study in their own right, but are beyond the scope of this modest coverage. Perhaps we will address these justifications a bit in lecture.

Either way, case-control data is shown in following 2-by-2 cross-tabulated form as follows:

	Disease+	**Disease-**	
Exposure +	a	b	n_1
Exposure -	c	d	n_0
	m_1	m_0	N

The *exposure proportion in cases* is:

$$\hat{p}_1 = a/m_1 \qquad \textbf{(14.1)}$$

The *exposure proportion in controls* is:

$$\hat{p}_0 = b/m_0 \qquad \textbf{(14.2)}$$

---

[1] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute, 11*, 1269-1275.

Neutra, R. R., & Drolette, M. E. (1978). Estimating exposure-specific disease rates from case-control studies using Bayes' theorem. *Am J Epidemiol, 108*(3), 214-222.

[2] Miettinen, O. (1976). Estimability and estimation in case-referent studies. *Am J Epidemiol, 103*(2), 226-235.

It makes intuitive sense that a higher exposure proportion in cases would justify an association between the exposure and disease. However, this relationship can be quantifies, we must convert the exposure proportions into exposure *odds*, as follows: let $\hat{q}_1 = 1 - \hat{p}_1 = c / m_1$ represent the exposure odds in cases and let $\hat{q}_0 = 1 - \hat{p}_0 = d / m_0$ represent the exposure odds in controls.

The *exposure odds ratio* is:

$$\hat{OR} = \frac{\hat{p}_1 / \hat{q}_1}{\hat{p}_0 / \hat{q}_0} = \frac{a / m_1 / c / m_1}{b / m_2 / d / m_0} = \frac{a/b}{c/d} = \frac{ad}{bc} \qquad \textbf{(14.3)}$$

It can be shown that this odds ratio is an estimate of the incidence (density) rate ratio which, itself, is an estimate of the risk ratio, especially when the disease is rare (risk < .05). Thus, we have another form of the relative risk. So much for theory.

Illustrative Example

As an illustrative example, let us consider a case-control study of esophageal cancer and alcohol consumption which included 200 cases and 775 controls.[3] Both cases and controls were administered a detailed dietary interview which contained questions about alcohol consumption, among many other factors. Data are contained in `BD1NEW.REC` as the variables `CASE` (1 = case, 2 = control) and `ALCHIGH` (alcohol consumption dichotomized at 80 grams per day: 1 = high, 2 = low). In tabular form, data are:

	Case	**Control**	
Exposure+	96	109	205
Exposure -	104	666	770
	200	775	975

From this we note $\hat{p}_1 = a / m_1 = 96 / 200 = 0.480$ and $\hat{p}_0 = b / m_0 = 109 / 775 = 0.141$. The odds ratio ($\hat{OR}$) = ad / bc = (96)(666) / (109)(104) = 5.640, suggesting that high-alcohol consumers have a much higher risk of esophageal cancer than light-consumers.

[3] Tuyns, 1977; Breslow & Day, 1980, Chapter 4

Confidence Interval for the *OR*

A 95% confidence interval for the *OR* is achieved by converting the point estimate of the odds ratio to a natural logarithm (ln) scale. For the illustrative example, $\ln \hat{O}R$ = ln(5.64) = 1.7299.

The standard error of ln odds ratio estimate is:

$$se_{\ln \hat{O}R} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \qquad \textbf{(14.4)}$$

For the illustrative data, $se_{\ln OR}$ = sqrt(1/96 + 1/109 + 1/104 + 1/666) = 0.1752.

A 95% confidence interval for the $\ln OR$ is now:

$$\ln \hat{O}R \pm (1.96)(se_{\ln \hat{O}R}) \qquad \textbf{(14.5)}$$

For the illustrative example, this confidence interval is 1.7299 ± (1.96)(0.1752) = 1.387, 2.073.

> Comment: The confidence level of the interval can be changed by altering the "1.96" to $z_{1-\alpha/2}$ as has been done elsewhere in *StatPrimer*.

To convert this confidence interval to **confidence interval for the *OR*,** take the anti-logarithm of these limits. For the illustrative example, a 95% confidence interval for $OR = e^{(1.387,\ 2.073)}$ = (4.0, 7.95).

Null Hypothesis Test

The test is traditionally done in a two-sided way. The null and alternative hypotheses are:

$$H_0: OR = 1 \text{ vs. } H_1: OR \underline{\text{not}} = 1$$

The test is performed with either a *chi-square test* or *Fisher's exact test*, depending on whether any expected frequencies are less than 5 (see prior chapter).

The expected frequencies for the illustrative data are:

	Disease+	**Disease-**	
Exposure +	42.051	162.949	205
Exposure -	157.949	612.0513	770
	200	775	975

Notice that all expected frequencies values exceed 5, so a chi-square method is appropriate.

The Yates' correct chi-square statistic in this instance (see Unit 16) is equal to $(|96 - 42.051| - .5)^2/42.051 + (|109 - 162.949| - .5)^2/162.949 + (|104 - 157.949| - .5)^2/157.949 + (|666 - 612.051| - .5)^2/612.051 = 108.22$ with 1 degree of freedom. The p value < .001; data are "significant." *qed*.

Z (Wald) Statistic

An alternative test statistic, based on the standard error estimate, is:

$$z_{\text{stat}} = \frac{\ln \hat{O}R}{se_{\ln \hat{O}R}} \qquad \textbf{(14.6)}$$

Under the null hypothesis, this statistic has a standard normal distribution.

For the illustrative example, $z_{\text{stat}} = 1.7299$, p (two-sided) < .002.

Matched (Paired) Samples

Introduction

Suppose we want to conduct a case-control study of disease D and exposure E in which cases and controls are uniquely-matched on potential confounders such as age, time, sex, clinic, and so on. We might do this to control for extraneous factors that could confound the study's results. In such general, such matched case-control pairs should be chosen to be alike in respected to all characteristics except for the exposure underinvestigation.

Although we might be tempted to use standard case-control procedures to analyze such data, methods that rely on sampling independence no longer apply: a new procedure is needed -- a procedure that accounts for the sample-match and addresses the fact that observations are no longer independent. This situation will now shift our attention from differences between individuals to differences *within* pairs, with data displayed as follows:

	Control pair-member is Exposed	**Control pair-member is Unexposed**
Case pair-member is E+	t	u
Case pair-member is E-	v	w

In this table cells t and w contain **concordant pairs** (case-control pairs-members are same with respect to exposure status), and cells u and v represent **discordant pairs** (case-control pairs-members are different with respect to exposure status). In analyzing these data, concordant pairs are ignored -- they provide little useful information about the potential effect of the exposure -- while we focus on the ratio of discordant pairs.

As an **illustrative example**, let us consider 50 age- and sex-matched case/controls-pairs.

	Control pair-member is E+	**Control pair-member is E-**
Case pair-member is E+	5	30
Case pair-member is E-	10	5

These data show 10 concordant pairs (cells a and d) and 30 discordant pairs (cells b and c). We will now discard the information from the concordant pairs (this may make you uncomfortable, but is the right thing to do), leaving an effective sample size of:

$$n' = u + v$$

For the illustrative example, $n' = 30 + 10 = 40$.

Confidence Interval for *OR*

An estimate for the odds ratio is provided by the ratio of the discordant pairs:

$$\hat{O}R = \frac{u}{v} \tag{14.7}$$

For the illustrative example, the odds ratio estimate = 30 / 10 = 3.0.

The standard error estimate of the ln(*OR*) is:

$$se_{\ln\hat{O}R} = \sqrt{\frac{1}{u} + \frac{1}{v}} \tag{14.8}$$

For the illustrative example, this standard error = sqrt(1/30 + 1/10) = 0.3651.

A 95% confidence interval for the ln*OR* is:

$$\ln(\hat{O}R) \pm 1.96(se_{\ln\hat{O}R}) \tag{14.9}$$

For the illustrative example, this = ln(3.0) ± 1.96(0.3651) = 1.0986 ± .7156 = (.3830, 1.8142).

To derive the confidence limits for the odds ratio, take the anti-log (base *e*) of these limits; a 95% confidence interval for the *OR* for the illustrative example = $e^{(.3830,\ 1.8142)}$ = (1.5, 6.2).

Null Hypothesis Test

The two-sided test of H_0: $OR = 1$ can be done with McNemar's chi-square statistic computed as:

$$c^2_{\text{McNemar, Uncorrected}} = \frac{(u-v)^2}{u+v} \qquad \textbf{(14.10)}$$

This statistic has 1 degree of freedom.

For the illustrative example, $\chi^2 = (30 - 10)^2 / (30 + 10) = 10.00$; $p < .01$.

The McNemar's test statistic can also be corrected for continuity, as follows:

$$c^2_{\text{McNemar, corrected}} = \frac{(|u-v|-1)^2}{u+v} \qquad \textbf{(14.11)}$$

This, too, is a chi-square statistic with 1 degree of freedom.

For the illustrative example, $\chi^2 = (|30 - 10| - 1)^2 / (30 + 10) = 9.025$; $p < .01$.

Z (Wald) Statistic

An alternative hypothesis testing statistic is:

$$z_{\text{stat}} = \frac{\ln \hat{O}R}{se_{\ln \hat{O}R}} \qquad \textbf{(14.12)}$$

For the illustrative example, $z_{\text{stat}} = 1.0986 / 0.3651 = 3.01$, p (two-sided) = .0026.