17: Case-Control Studies (Odds Ratios)

Independent Samples

The prior chapter use risk ratios from cohort studies to quantify exposure–disease relationships. This chapter uses odds ratios from case-control studies for the same purpose.

We will discuss the sampling theory behind case-control studies in lecture. For details, see pp. 208– 212 in my text *Epidemiology Kept Simple*.

The general idea is to select all cases in the population and a simple random sample of non-cases (controls). The cross-tabulated data looks like this:

Exposure variable	**Response variable** +	−	Total
+	a_1	b_1	n_1
−	a_2	b_2	n_2
Total	m_1	m_2	N

Case-control studies can *not* calculate incidences or prevalences. They can, however, calculate exposure **odds ratios**:

$$\hat{O}R = \frac{A_1 B_2}{A_2 B_1}$$

This statistic, which is just the **cross-product ratio** of the entries in the 2-by-2 table, is an estimate of the relative incidence (*relative risk*) of the outcome associated with exposure (assuming data are error-free).

The **confidence interval for the *OR*** parameter is

$$e^{\ln \hat{O}R \pm z \cdot SE_{\ln \hat{O}R}}$$

where e is the base on the natural logarithms ($e \approx 2.71828\ldots$), z is a Standard Normal deviate corresponding to the level of confidence ($z = 1.645$ for 90% confidence, $z = 1.96$ for 95% confidence, and $z = 2.576$ for 99% confidence), and

$$SE_{\ln \hat{O}R} = \sqrt{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{b_1} + \frac{1}{b_2}}\ .$$

A **test** of H_0: $OR = 1$ is calculated with a chi-square statistic or Fisher's test, depending on the size of the sample (see prior chapter).

Example:* *Alcohol and esophageal cancer. Data from a case-control study of 200 esophageal cancer cases and 775 community-based controls are shown below.[1] Detailed dietary data were obtained by interview. This example addresses the relation between alcohol consumption (dichotomized at 80 grams per day) and esophageal cancer. Data are:

Alcohol g/day	**Esophageal cancer** +	−	Total
+	96	109	205
−	104	666	770
Total	200	775	975

The odds ratio = (96)(666)/(109)(104) = 5.6401 = 5.64, suggesting esophageal cancer is 5.64 times as frequent in the exposed group in the source population.

To calculate confidence intervals, note that ln(ψ^) = ln(5.640) = 1.7299 (by calculator) and standard error $SE_{\ln \hat{OR}} = \sqrt{\frac{1}{a_1}+\frac{1}{a_2}+\frac{1}{b_1}+\frac{1}{b_2}} = \sqrt{\frac{1}{96}+\frac{1}{104}+\frac{1}{109}+\frac{1}{665}}$ = 0.1752. The 95% confidence interval for the ψ $= e^{1.7299 \pm (1.96)(0.1752)} = e^{1.7299 \pm 0.3433} = e^{1.3866,\ 2.0732}$ = 4.00 to 7.95. The 90% confidence interval for the ψ $= e^{1.7299 \pm (1.645)(0.1752)} = e^{1.7299 \pm 0.2882} = e^{1.4417,\ 2.0181} = e^{1.4417,\ 2.0181}$ = 4.23 to 7.52.

The *P*-value for testing H_0: ψ = 1 can be derived by chi-square test. In this case, X^2_{stat}= 110.26 and $X^2_{stat,\ cont\text{-}corrected}$ = 108.22. Both have 1 *df* and both derive *P* ≈ 0.00000.

Results may be confirmed with SPSS (individual records), WinPepi or EpiCalc2000 (cross-tabulated data).

As always, the primary threats in practice are systematic errors (bias), not random, errors (imprecision).

Matched samples

A matched design may be used in both cohort and case-control studies to help control for confounding by extraneous factors.

For cohort data, matched-pairs are displayed as follows:

Exposed pair-member	**Non-exposed pair-member** Case	Non-case	Total
Case	t	u	n_1
Non-case	v	w	n_2
Total	m_1	m_2	N

For case-control data, matched-pairs are displayed as follows:

Case pair-member	**Control pair-member** Exposed	Non-exposed	Total
Exposed	t	u	n_1
Non-exposed	v	w	n_2
Total	m_1	m_2	N

Counts in this table represent the numbers of pairs, not numbers of individuals. Cells t and w in this table contain the number of **concordant pairs** in the sample. Concordant pairs are the same with respect to exposure. Cells u and v contain **discordant pairs**. *Discordant* pairs differ with respect to exposure. Although there are N pairs total, we are interested only in the $(u + v)$ discordant pairs.

The **odds ratio** for these data is:

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

The **confidence interval for** ψ is

$$e^{\ln \hat{O}R \pm z \cdot SE_{\ln \hat{O}R}}$$

where e is the base on the natural logarithms ($e \approx 2.71828\ldots$), z is a Standard Normal deviate corresponding to the desired level of confidence ($z = 1.645$ for 90% confidence, $z = 1.96$ for 95% confidence, and $z = 2.576$ for 99% confidence), and $SE_{\ln \hat{O}R} = \sqrt{\frac{1}{u} + \frac{1}{v}}$.

When the number of discordant pairs $(u + v)$ is 10 or greater, you can test H_0: $OR = 1$ with McNemar's chi-square statistic. The regular and continuity-correct McNemar's chi-squares are shown below:

$$X^2_{\text{McN}} = \frac{(u-v)^2}{u+v}$$

$$X^2_{\text{McN,cc}} = \frac{(|u-v|-1)^2}{u+v}$$

McNemar's chi-square statistics have 1 df.

Because of the relation between the chi-square distributions and z distributions, the above formulas can be re-expressed:

$$z_{\text{stat, McN}} = \sqrt{\frac{(u-v)^2}{u+v}}$$

$$z_{\text{stat, McN cc}} = \sqrt{\frac{(|u-v|-1)^2}{u+v}}$$

With small samples, let the number of positive discordant pairs (u) be the numerator of a proportion and let the total number of discordant pairs ($u + v$) represent the denominator of a proportion. Then test, H_0: $p = ½$ with an exact binomial test (see Chapter 16 in the new biostat-text for details).

Example. ***Matched cohort data (Smoking and mortality in identical twins)*.** When smoking was first suspected as a cause of disease, Sir Ronald Fisher offered the *constitution hypothesis* as an alternative explanation for the observed association. The constitutional hypothesis suggested that people genetically disposed to lung cancer were more likely to smoke. In other words, the relation between smoking and disease was *confounded* by constitutional factors. The constitutional hypothesis was put to the ultimate test by a study in which 22 smoking-discordant monozygotic twins where studied to see which twin first succumbed to death.[2] In this study, the smoking-twin died first in 17 of the pairs (i.e., $u = 17$, $u + v = 22$, so $v = 5$).

The odds ratio estimate $\hat{OR} = \frac{u}{v} = \frac{17}{5} = 3.40$. The smoking twin was 3.4 as likely to die first.

In testing, H_0: $OR = 1$, $z_{\text{stat, McN}} = \sqrt{\frac{(u-v)^2}{u+v}} = \sqrt{\frac{(17-5)^2}{17+5}} = 2.56$; $P = 0.010$. With continuity correction, $z_{\text{stat, McN cc}} = \sqrt{\frac{(|u-v|-1)^2}{u+v}} = \sqrt{\frac{(|17-5|-1)^2}{17+5}} = 2.35$; $P = 0.019$), providing "significant" evidence against the null hypothesis. Thus the constitutional hypothesis is refuted and for the causal hypothesis is supported.

(This example illustrates how statistical testing can be used as a small part of dealing with the uncertainty connected with scientific inference.)

Example. ***Matched case-control data (Fruits, vegetables, and adenomatous polyps)*****.** A case-control study used matched-pairs to study the statistical relationship between adenomatous polyps of the colon in relation to diet. Cases and controls in the study had undergone sigmoidoscopic screening. Controls were matched to cases on time of screening, clinic, age, and sex. One of the study's statistical analyses considered the effects of low fruit and vegetable consumption on colon polyp risk. There were 45 pairs in which the case but not the control reported low fruit/veggie consumption. There were 24 pairs in which the control but not the case reported low fruit/veggie consumption.[3]

Based on this information, the odds ratio estimate $\hat{O}R = \frac{u}{v} = \frac{45}{24}$ = 1.88, indicating that low fruit/veggie "exposure" was associated with an 88% increase in risk.

The 95% confidence interval for the odd ratio parameter is calculated. The ln(*OR*^) = 0.6286 and $SE_{\ln \hat{O}R} = \sqrt{\frac{1}{u}+\frac{1}{v}} = \sqrt{\frac{1}{45}+\frac{1}{24}}$ = 0.2528. Therefore, the 95% confidence interval for $OR = e^{0.6286 \pm (1.96)(0.2528)} = e^{0.6286 \pm 0.4959} = \Psi = e^{(0.1331,\ 1.1241)}$ = (1.14, 3.07)

In testing, H_0: *OR* = 1, $z_{\text{stat, McN}} = \sqrt{\frac{(u-v)^2}{u+v}} = \sqrt{\frac{(45-24)^2}{45+24}}$ = 2.53; *P* = 0.011. With continuity correction, $z_{\text{stat, McN cc}} = \sqrt{\frac{(|u-v|-1)^2}{u+v}} = \sqrt{\frac{(|45-24|-1)^2}{45+24}}$ = 2.41; *P* = 0.016.

References

[1] Tuyns, A. J., Pequignot, G., & Jensen, O. M. (1977). [Esophageal cancer in Ille-et-Vilaine in relation to levels of alcohol and tobacco consumption. Risks are multiplying]. *Bulletin du Cancer, 64*(1), 45-60.
[2] Kaprio, J., & Koskenvuo, M. (1989). Twins, smoking and mortality: a 12-year prospective study of smoking- discordant twin pairs. *Social Science & Medicine, 29*(9), 1083-1089.
[3] Witte, J. S., Longnecker, M. P., Bird, C. L., Lee, E. R., Frankl, H. D., & Haile, R. W. (1996). Relation of vegetable, fruit, and grain consumption to colorectal adenomatous polyps. *American Journal of Epidemiology, 144*(11), 1015-1025. Summary of frequencies reported in Rothman & Greenland, 1998, p. 287.