CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 13

Estimating Relative Risks from Case-Control Data and Sensitivity Analyses

Motivating example. Our motivating example for this lecture comes from Chapter 5 of "Statistics in Public Health" (1998) edited by Donna Stroup and Steven Teutsch.

On pages 102 and 103 the authors describe a fictional case-control study: 100 patients who have *E. coli* infection and 100 controls were asked if they drank a particular brand of apple juice within the last week.

The authors then present three tables of fictional data, reproduced below as Tables 1 through 3. The authors argue that, with data as in Table 1, there is no doubt of an association between apple juice consumption and *E. coli* infection. On the other hand, with data as in Tables 2 and 3, a formal statistical analysis is necessary to resolve whether there is such an association.

TABLE 1

Sample	Yes (Drank)	No (Did Not Drink)	Row Total
Cases	95	5	100
Controls	5	95	100
Column Total	100	100	200

TABLE 2

Sample	Yes (Drank)	No (Did Not Drink)	Row Total
Cases	60	40	100
Controls	45	55	100
Column Total	105	95	200

TABLE 3

Sample	Yes (Drank)	No (Did Not Drink)	Row Total
Cases	54	46	100
Controls	40	60	100
Column Total	94	106	200

An astonishing assertion. In describing Table 2, the authors assert that "the risk for disease among those who drank can be estimated as $60/(60+45) = 0.5714$, and, among those who did not, as $40/(40+55) = 0.4211$. The ratio of estimated risks, called the approximate relative risk, is $0.5714/0.4211 = 1.357$."

Discussion Questions. If the *E. coli* outbreak occurred in Lexington and the authors' estimates of the risks were close to the truth, then how many people in Lexington would be sick? Why is such a scenario not believable? What was the authors' conceptual mistake?

Relative risk and odds ratio. Let p_1 denote the risk of *E. coli* infection for someone who drinks the brand of apple juice in question. Also, let $O_1 = p_1/(1-p_1)$ denote the odds of *E. coli* infection for someone who drinks the apple juice.

Let p_2 denote the risk of *E. coli* infection for someone who does not drink the apple juice. Also, let $O_2 = p_2/(1-p_2)$ denote the odds of *E. coli* infection for someone who does not drink the apple juice.

From CPH 930 you know that the relative risk of *E. coli* infection is defined by

$$p_1/p_2,$$

while the odds ratio is given by

$$O_1/O_2 = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

Estimating the relative risk. Since the numbers 0.5714 and 0.4211 grossly overstate the risks p_1 and p_2, there is no reason to believe that $0.5714/0.4211 = 1.357$ should be a good estimate of the relative risk p_1/p_2.

This raises the question of how we can estimate the relative risk from case-control data. In CPH 930 you were given one option: estimate the odds ratio instead and then assume that the estimated odds ratio could serve as an estimate of the relative risk.

In CPH 931 we now pursue another option. Let E denote the event that a person drinks the apple juice, and let D denote the event that a person has the *E. coli* infection.

By Bayes' Theorem, we have

$$p_1 = P(D|E) = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|\bar{D})P(\bar{D})}$$

and

$$p_2 = P(D|\bar{E}) = \frac{P(\bar{E}|D)P(D)}{P(\bar{E}|D)P(D) + P(\bar{E}|\bar{D})P(\bar{D})}.$$

Since the 100 cases may be viewed as a random sample of those infected with *E. coli*, we may reasonably regard $60/100 = 0.60$ as an estimate of

$P(E|D)$. Likewise, we may reasonably regard $45/100 = 0.45$ as an estimate of $P(E|\bar{D})$. Hence,

$$p_1 = P(D|E) \approx \frac{0.60P(D)}{0.60P(D) + 0.45P(\bar{D})} = \frac{0.60P(D)}{0.15P(D) + 0.45}$$

and

$$p_2 = P(D|\bar{E}) \approx \frac{0.40P(D)}{0.40P(D) + 0.55P(\bar{D})} = \frac{0.40P(D)}{0.55 - 0.15P(D)}.$$

Consequently,

$$p_1/p_2 \approx 1.5 \times \frac{0.55 - 0.15P(D)}{0.15P(D) + 0.45}.$$

All we need to estimate the relative risk is an estimate of $P(D)$, the prevalence of *E. coli* infection.

Note that, as the prevalence $P(D)$ tends to 1, both risk estimates tend to 1 and so the estimate of the relative risk tends to 1. On the other hand, as the prevalence $P(D)$ tends to 0, both risk estimates tend to 0. You know from calculus that a limit of the form $0/0$ can turn out to be virtually anything, and in this case the limit turns out to be

$$1.5 \times \frac{0.55}{0.45} = 1.833.$$

We will see shortly what the interpretation of this limit is. Before doing that, let us see what happens with a couple different values of $P(D)$.

Discussion Questions. What is the estimate of the relative risk if $P(D) = 0.001$? What if $P(D) = 0.5$? Does the latter number look familiar?

Estimating the odds ratio. The above approach to estimating the relative risk in a case-control study requires an estimate of the prevalence. However, we can estimate the odds ratio without an estimate of the prevalence.

To see why, note that

$$1-p_1 = 1-\frac{P(E|D)P(D)}{P(E|D)P(D)+P(E|\bar{D})P(\bar{D})} = \frac{P(E|\bar{D})P(\bar{D})}{P(E|D)P(D)+P(E|\bar{D})P(\bar{D})}$$

and

$$1-p_2 = 1-\frac{P(\bar{E}|D)P(D)}{P(\bar{E}|D)P(D)+P(\bar{E}|\bar{D})P(\bar{D})} = \frac{P(\bar{E}|\bar{D})P(\bar{D})}{P(\bar{E}|D)P(D)+P(\bar{E}|\bar{D})P(\bar{D})}.$$

The above computations yield

$$O_1 = \frac{p_1}{1-p_1} = \frac{P(E|D)P(D)}{P(E|\bar{D})P(\bar{D})} \quad \text{and} \quad O_2 = \frac{p_2}{1-p_2} = \frac{P(\bar{E}|D)P(D)}{P(\bar{E}|\bar{D})P(\bar{D})}.$$

When we divide O_1 by O_2, the factors of $P(D)$ and $P(\bar{D})$ cancel, leaving us with

$$O_1/O_2 = \frac{P(E|D)\times P(\bar{E}|\bar{D})}{P(E|\bar{D})\times P(\bar{E}|D)}.$$

But $P(E|D)$, $P(\bar{E}|\bar{D})$, $P(E|\bar{D})$, and $P(\bar{E}|D)$ can all be estimated directly from the contingency table in a case-control study.

Using the data from Table 2, for example, we have

$$O_1/O_2 \approx \frac{0.60\times 0.55}{0.45\times 0.40} = 1.833.$$

This computation suggests, and it is in fact true that, the estimated relative risk in a case-control study converges to the estimated odds ratio as the estimated prevalence tends to 0.

This computation also illustrates the familiar formula $\frac{a\times d}{b\times c}$, where a through d are the four entries in the two-by-two contingency table summarizing the results from a case-control study.

Sample sizes in case-control studies. As noted on page 98, case-control studies are "extremely useful for investigating outbreaks in a community, especially when the disease is rare ... and when results are needed quickly."

Although the oversampling of cases impedes estimation of p_1 and p_2 individually, it is precisely this feature of case-control studies that permits assessment of the association between disease and exposure with only a few dozen or a few hundred study participants rather than a few thousand.

To better understand this idea, recall from CPH 930 that the 95% confidence interval for an odds ratio — whether obtained from a prospective cohort study or a retrospective case-control study — is given by

$$\frac{a \times d}{b \times c} \; \exp\left[\pm 1.96\sqrt{1/a + 1/b + 1/c + 1/d}\right].$$

For instance, the 95% confidence interval based on the data in Table 2 is

$$\begin{aligned} 1.833 \exp&\left[\pm 1.96\sqrt{1/60 + 1/40 + 1/45 + 1/55}\right] \\ &= 1.833 \exp\left[\pm 1.96(0.2865)\right] \\ &= 1.046 \text{ to } 3.214. \end{aligned}$$

Generally speaking, we conclude that an estimated odds ratio $\widehat{OR}$ is significantly greater than 1 — and that there is an association between disease and exposure — when the lower bound for the 95% confidence interval exceeds 1. This is true if and only if

$$\widehat{OR} > \exp\left[1.96\sqrt{1/a + 1/b + 1/c + 1/d}\right].$$

Rearranging the inequality yields

$$1/a + 1/b + 1/c + 1/d < \left(\log\left[\widehat{OR}\right]/1.96\right)^2.$$

Since $1/b$, $1/c$, and $1/d$ must be positive, a necessary but not sufficient condition for the above inequality to hold is that

$$1/a < \left(\log\left[\widehat{OR}\right]/1.96\right)^2,$$

which is equivalent to

$$a > \left(1.96/\log\left[\widehat{OR}\right]\right)^2.$$

<u>Thus, we have no hope of establishing an association between disease and exposure if the number of cases with the exposure is less than the square of 1.96 divided by the logarithm of the estimated odds ratio.</u>

For instance, suppose that $\widehat{OR} = 1.833$. Then

$$\left(1.96/\log\left[\widehat{OR}\right]\right)^2 = (1.96/0.606)^2 = 10.46.$$

The 95% confidence interval for the odds ratio is mathematically guaranteed to contain 1 unless $a \geq 11$, regardless of how large b or c or d may be.

What does this have to do with study design? Suppose that $P(D|E)$ is a small number like 0.01 but that $P(E|D)$ is a larger number like 0.50. In words, 1% of exposed individuals have the disease, but 50% of diseased individuals have the exposure.

To achieve $a = 11$ with a prospective cohort study will require us to sample approximately $11/0.01 = 1100$ exposed individuals, but to achieve $a = 11$ with a retrospective case-control study will require us to sample approximately $11/0.50 = 22$ diseased individuals. Clearly, the latter sampling scheme is more economical.

Remarks. Several comments can be made on the preceding developments.

We can replace a above by b, c, or d to conclude that the following four conditions are collectively necessary but not sufficient for the 95% confidence interval to exclude 1. In other words, the confidence interval is guaranteed to include 1 when any one of these conditions fails. However, the confidence interval may or may not exclude 1 when all four conditions are met.

1. $a > \left(1.96/\log\left[\widehat{OR}\right]\right)^2$

2. $b > \left(1.96/\log\left[\widehat{OR}\right]\right)^2$

3. $c > \left(1.96/\log\left[\widehat{OR}\right]\right)^2$

4. $d > \left(1.96/\log\left[\widehat{OR}\right]\right)^2$

On the other hand, the next four conditions are collectively sufficient for the 95% confidence interval to exclude 1. That is, the confidence interval is guaranteed to exclude 1 when all four of these conditions are met.

1. $a > 4\left(1.96/\log\left[\widehat{OR}\right]\right)^2$

2. $b > 4\left(1.96/\log\left[\widehat{OR}\right]\right)^2$

3. $c > 4\left(1.96/\log\left[\widehat{OR}\right]\right)^2$

4. $d > 4\left(1.96/\log\left[\widehat{OR}\right]\right)^2$

Thus, with $\widehat{OR} = 1.833$, we are guaranteed a 95% confidence interval that excludes 1 if a, b, c, and d are all at least $42 > 41.84 = 4 \times 10.46$.

Of course, $\widehat{OR}$ is itself a function of a, b, c, and d. This has two implications. First, we have to use a projected value for $\widehat{OR}$ in the inequalities above, much as we have to use projected values for means and standard deviations in the sample size formulas from an introductory statistics course. Second, there is a constraint on a, b, c, and d besides their being at least 42. Having $a = b = c = d = 42$ will not work because then $\widehat{OR} = 1$ and obviously the 95% confidence interval will contain 1. What we can do, however, is keep recruiting cases until $\min\{a, b\} \geq 42$ and keep recruiting controls until $\min\{c, d\} \geq 42$. Note that two out of a, b, c, and d may end up being greater than 42, possibly much greater.

Sensitivity analysis. Sensitivity analysis can take many different forms. The general idea is to determine how sensitive the end product of a statistical analysis is to some (possibly) controversial decision that the investigator must make in carrying out the analysis. Here are some examples.

1. We can see how much the end product of a statistical analysis changes when an input parameter is varied. For instance, the end product may be an estimated relative risk for *E. coli* infection, while the input parameter may be our guess for the prevalence of *E. coli* infection.

2. We can see how much the end product of a statistical analysis changes when a certain influential or outlying observation is removed. For instance, the end product may be a parameter estimate in a linear regression model, while the outlying observation could be one that we have flagged due to a large externally studentized residual. In fact, the DFBETA that you learned about in CPH 930 <u>is</u> a measure of how much a parameter estimate changes when an observation is removed, scaled in such a way that $\sqrt{n}$ times the DFBETA can be judged in relation to a standard normal distribution.

3. We can see how much the end product of a statistical analysis changes when the estimation procedure is modified. For instance, the end product may be a parameter estimate in a linear regression model, while the estimation procedure may be ordinary least squares or one of the other (parametric) estimation procedures from Lectures 1 and 2.

4. We can see how much the end product of a statistical analysis changes when a distributional assumption is modified. For instance, the end product may be a rate ratio estimate in a generalized linear model, while the distributional assumption for the total number of events in a homogeneous subgroup may be Poisson or negative binomial.

A sensitivity analysis for our motivating example. Let c denote the prevalence of *E. coli* infection. Define

$$p_1(c) := \frac{P(E|D) \times c}{P(E|D) \times c + P(E|\bar{D}) \times (1-c)}$$

and

$$p_2(c) := \frac{P(\bar{E}|D) \times c}{P(\bar{E}|D) \times c + P(\bar{E}|\bar{D}) \times (1-c)},$$

which represent the risks of *E. coli* infection for apple juice drinkers and non apple juice drinkers as functions of the prevalence.

If we estimate $P(E|D)$ and $P(E|\bar{D})$ by 0.60 and 0.45 using the data from Table 2, then we obtain

$$\widehat{p}_1(c) := \frac{0.60 \times c}{0.60 \times c + 0.45 \times (1-c)}$$

and

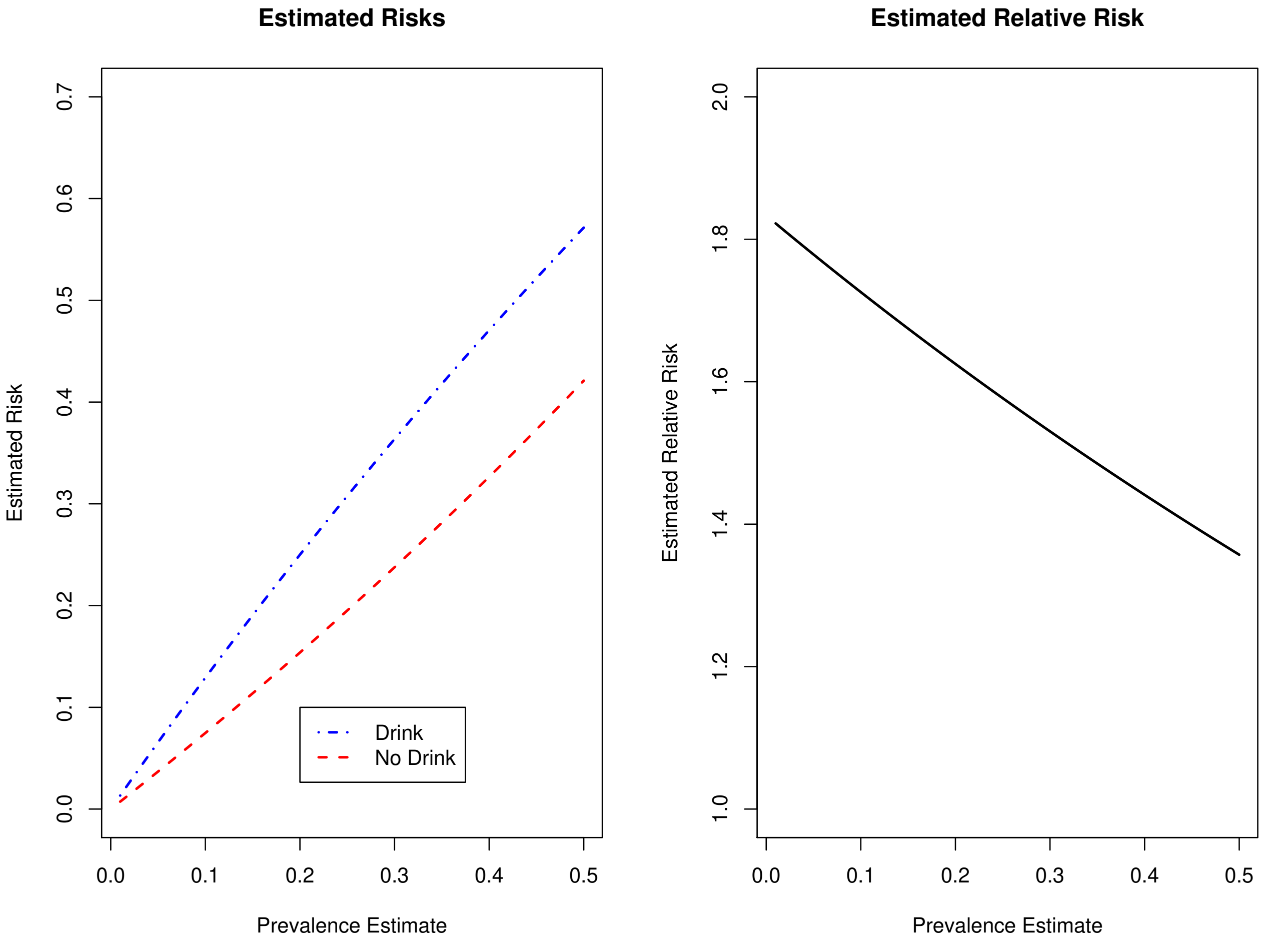
$$\widehat{p}_2(c) := \frac{0.40 \times c}{0.40 \times c + 0.55 \times (1-c)}.$$

The left panel of Figure 1 shows $\widehat{p}_1(c)$ (blue alternating short and long dashes) and $\widehat{p}_2(c)$ (red long dashes) as functions of c, while the right panel of Figure 1 shows $\widehat{p}_1(c)/\widehat{p}_2(c)$ (black solid).

As the prevalence increases, the estimated risks of *E. coli* infection increase for both apple juice drinkers and non apple juice drinkers. However, the estimated relative risk decreases!

Note that the authors' estimates are recovered at $c = 0.50$. Since a prevalence as high as 0.50 is dubious, the authors not only overstate the risks of *E. coli* infection but also appear to understate the relative risk: to the extent that we can trust the point estimate, the association between *E. coli* infection and apple juice consumption is much stronger than the authors asserted!

FIGURE 1


Estimated Risks
Estimated Risk
Prevalence Estimate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Drink
No Drink
Estimated Relative Risk
Estimated Relative Risk
Prevalence Estimate
1.0
1.2
1.4
1.6
1.8
2.0