February 2003: The Odds Ratio Approximates the Relative Risk Assuming that the Disease is Rare (Rule 4.2)

Introduction

Rules of the month are numbered in accordance with the numbering in the book. Thus, Rule 1.1 refers to the first rule in Chapter 1. And so on. These comments do not repeat the material in the book but highlights and amplifies it. A rule is stated as found in the book and then discussed.

Rule 4.2 "Under the rare-disease assumption the odds ratio approximates the relative risk."

Further Comments on the Rule

The discussion in this section of the book emphasized three kinds of studies: cross-sectional, retrospective, and prospective. The prerequisite for some kind of random sampling was perhaps not stressed enough. In this note I want to elaborate on the importance of random sampling of the populations.

Suppose the population is divided into the following four categories, as in Table 4.3 in the book,

	Disease (D+)	No Disease (D-)	Total
Exposed (E+)	π_{11}	π_{12}	$\pi_{1\bullet}$
Not Exposed (E-)	π_{21}	π_{22}	$\pi_{2\bullet}$
Total	$\pi_{\bullet 1}$	$\pi_{\bullet 2}$	1

This formulation emphasizes that the total population is split into these four categories. Suppose we now randomly select arbitrary proportions from each of these four categories. That is, we select a proportion p_{11} of subjects from the exposed and diseased populations. Similarly, we select proportions p_{ij} from the other populations. Then the sample proportions will be,

	Disease (D+)	No Disease (D-)	Total
Exposed (E+)	$p_{11}\pi_{11}$	$p_{12}\pi_{12}$	
Not Exposed (E-)	$p_{21}\pi_{21}$	$p_{22}\pi_{22}$	
Total			

I have left off the marginal totals because their interpretation will depend on the sampling scheme. At this point we simple assume an arbitrary sampling scheme for each of the four cells (although still assumed to be random from that cell's population).
First, calculate the relative risk from this sample (whatever that number may mean),

$$R = \frac{p_{11}\pi_{11} \Big/ (p_{11}\pi_{11} + p_{12}\pi_{12})}{p_{21}\pi_{21} \Big/ (p_{21}\pi_{21} + p_{22}\pi_{22})}.$$

Under what sampling conditions is this ratio equal to the relative risk? A trivial situation is, $p_{11}=p_{12}=p_{21}=p_{22}=p$, which is equivalent to cross-sectional sampling, that is, all the cells are sampled equally. Another case where this is true is where $p_{11}=p_{12}$ and $p_{21}=p_{22}$; essentially cohort sampling.

Are there other sampling schemes where the population relative risk can still be estimated? Yes, when the sampling odds ratio is 1! That is, $p_{11}p_{22}=p_{12}p_{21}$ or $p_{11}p_{22}-p_{12}p_{21}=0$. With small deviations from 1(or 0, in the case of the differences) the bias will be small. This scenario is not too helpful since the sample odds ratio is precisely what we use to estimate the odds ratio! However, this point is made to indicate that sampling does not have to be simple random sampling.

A similar argument can be made about the odds ratio. Under the arbitrary sampling scheme defined above the odds ratio is,

$$O = \frac{p_{11}p_{22}}{p_{12}p_{21}} \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Under all the sampling schemes discussed in the previous paragraph the odds ratio will be estimated without bias. There is, of course, the additional sampling scheme of case-control that also eliminates the bias. That is, $p_{11}=p_{21}$ and $p_{12}=p_{22}$, corresponding to selecting a specified proportion from the diseased population and classifying these cases with respect to exposure status, and another proportion from the control population and classifying these controls with respect to exposure status.

Now switch gears slightly by assuming that the row categories are not exposure status but outcomes of a screening test, and we want to test sensitivity and specificity. The sensitivity is defined to be,

$$\text{Sensitivity} = \text{Sn} = \frac{p_{11}\pi_{11}}{p_{11}\pi_{11} + p_{21}\pi_{21}} = \frac{\pi_{11}}{\pi_{11} + \frac{p_{21}}{p_{11}}\pi_{21}}.$$

The first equation indicates that the sensitivity is only estimated in an unbiased manner if $p_{11}=p_{21}$, that is, the equivalent assumption of case-control sampling. The second equation provides scenarios of what happens when the sampling is not equal. For example, if $p_{21}<p_{11}$, the sensitivity will be overestimated. But does this ever happen in practice? My guess is that it's more common than is suspected. In cognitive testing in Alzheimer's disease there are subtle selection biases. For example, better educated subjects usually appear in clinics. If such subjects are less likely to score negative on the screen then the sensitivity of the test will be exaggerated.

A similar scenario can be considered with respect to specificity.

$$\text{Specificity} = \text{Sp} = \frac{p_{22}\pi_{22}}{p_{22}\pi_{22} + p_{12}\pi_{12}} = \frac{\pi_{22}}{\pi_{22} + \frac{p_{12}}{p_{22}}\pi_{12}}.$$

Again, if $p_{12}<p_{22}$ the specificity will be biased upward. This will happen in cognitive testing where better educated people will tend to score negative on a test and better educated people are more likely to volunteer to be controls in a screening situation.

I suspect someone has done this kind of modeling before. If anyone reading this can give me a reference, I would appreciate that and will give the appropriate acknowledgments; it's likely that it's been done better as well.